REACT Fits to Linear Models and Scatterplots

Rudolf Beran*
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720–3860, USA

September 1998

REACT estimators for the mean of a linear model involve three steps: transforming the model to a canonical form that provides an economical representation of the unknown mean vector, estimating the risks of a class of candidate linear shrinkage estimators, and adaptively selecting the candidate estimator that minimizes estimated risk. When the mean vector is smooth, the desired canonical form of the linear model is achieved by constructing a smooth orthogonal basis for the regression space. Such a smooth basis for a complete, balanced one-way layout is asymptotically equivalent to the discrete cosine basis. Applied to one- or higher-way layouts, the REACT method generates automatic scatterplot smoothers that compete well on standard data sets with the best fits obtained by alternative techniques. Historical precursors to REACT include nested model selection, ridge regression, and nested principal component selection for the linear model. However, REACT's insistence on working with an economical basis greatly increases its superefficiency relative to the least squares fit. A secondary improvement stems from REACT's use of flexible monotone shrinkage rather than 0-1 shrinkage of components. Both improvements are demonstrated numerically on data sets and theoretically through Pinsker bounds for minimax risk in the estimation problem.

AMS classification: 62J05, 62G07
Keywords and phrases: risk estimation, adaptation, discrete cosine transform, economical basis, minimum C_L, symmetric linear smoother, asymptotic minimax, shrinkage.

1. INTRODUCTION

The acronym REACT stands for **r**isk **e**stimation, **a**daptation, **c**oordinate **t**ransformation. These are the three components of a methodology, described in this paper, that yields superefficient fits to the Gaussian linear model. The risk of REACT fits under quadratic loss is often far smaller than the risk of the classically efficient least squares fit. Applied to the one-way layout, REACT fits generate automatic scatterplot smoothers that compete well

* Research supported in part by National Science Foundation Grant DMS95–30492 and, at Universität Heidelberg, by the Alexander von Humboldt Foundation.

on test data with kernel or local polynomial techniques. Similarly, REACT for the two-way layout provides an effective technique for fitting response surfaces to observations taken over a two-dimensional grid.

Consider a Gaussian linear model in which the $n \times 1$ response vector y has a $N_n(X\beta, \sigma^2 I_n)$ distribution. For simplicity, assume that the $n \times p$ regression matrix X has full rank $p \leq n$. Both the regression coefficients β and the variance σ^2 are unknown. The problem is to estimate $\eta = \mathrm{E}(y) = X\beta$. The risk of any estimator $\hat{\eta}$ is the expectation of the normalized quadratic loss $p^{-1}|\hat{\eta} - \eta|^2$. In particular, the risk of the classical least squares estimator $\hat{\eta}_{LS} = X(X'X)^{-1}X'y$ is σ^2. Stein (1956) proved that $\hat{\eta}_{LS}$ is inadmissible whenever the dimension p of the regression space exceeds 2. While the theoretical depth of his result was recognized quickly, development of its extensive implications for statistical practice has been slow. However, the essential flaw in least squares—its propensity to overfit a linear model when p is not small—has motivated work on principal component regression, ridge regression, and model selection.

We outline the main steps of the REACT methodology:

1) *Coordinate transformation.* By orthogonal transformation of y, reduce the model to standard canonical form: z and $\bar{z}$ are independent, z has $N_p(\xi, \sigma^2 I_p)$ distribution, and $\bar{z}$ has $N_{n-p}(0, \sigma^2 I_{n-p})$ distribution. Sensible choice of the orthogonal transformation is essential in obtaining REACT fits with small risk. Ideally, all but the first few components of ξ should be nearly zero.

2) *Risk estimation.* Let $\mathcal{F}$ be a subset of $[0,1]^p$ such that the components of each vector $f \in \mathcal{F}$ are monotone nonincreasing. Consider candidate estimators of ξ of the form $\hat{\xi}(f) = fz$. The multiplication here is componentwise, as in S code. Such an estimator $\hat{\xi}(f)$ shrinks the components of z, which is the least squares estimator of ξ, downscaling especially the higher-order components of z. For every $f \in \mathcal{F}$, estimate the risk of $\hat{\xi}(f)$ from the data.

3) *Adaptation.* Find $\hat{f} \in \mathcal{F}$ that minimizes the estimated risk function from step 2. Estimate ξ by $\hat{\xi}(\hat{f})$. Mapping this adaptive estimator of ξ back into an estimator of η yields the REACT fit.

Scatterplot fits can be related to the one-way layout through the following model: given $\{x_i : 1 \leq i \leq n\}$, suppose that $\{y_i : 1 \leq i \leq n\}$ are conditionally independent and that the conditional distribution of y_i is $N(m(x_i), \sigma^2)$ for every i. If the function m is unknown and no ties exist among the $\{x_i\}$, then this conditional scatterplot model is equivalent to the one-way layout with one observation per cell—the linear model with $X = I_n$. When there are ties among the $\{x_i\}$, the scatterplot model is equivalent to an unbalanced one-way layout, a linear model in which each row of the regression matrix X has a single nonzero entry that takes the value 1; rows are repeated to reflect the pattern of ties among the $\{m(x_i)\}$. By reordering labels as necessary, suppose that x_i is nondecreasing as a function of i. Then

linear (or other) interpolation between the successive components of the REACT estimator for the conditional means produces a curve fit to the scatterplot.

Section 2 describes REACT fits in detail. Their application to one-way layouts is illustrated on scatterplots drawn from the smoothing literature. Heteroscedasticity, two-way layouts, and confidence sets centered at REACT fits are among the subjects of Section 3. Section 4 presents theoretical properties of REACT fits, relating these to estimators that achieve the Pinsker (1980) asymptotic minimax bound.

2. PROCEDURE AND EXAMPLES

To completely define REACT fits requires specifying the orthogonal transformation in step 1, the class of candidate estimators and risk estimator in step 2, and the computational algorithm for the minimization in step 3. We consider these matters in turn.

2.1 Choice of orthonormal basis

For any matrix A, let $\mathcal{M}(A)$ denote the subspace spanned by the columns of A. Let U be an $n \times p$ matrix with orthonormal columns such that $\mathcal{M}(U) = \mathcal{M}(X)$. Select the $n \times (n-p)$ matrix $\bar{U}$ so that $O = (U|\bar{U})$ is an orthogonal matrix and define

$$z = U'y, \qquad \bar{z} = \bar{U}'y, \qquad \xi = \mathrm{E}z = U'\eta. \tag{1}$$

Such orthogonal transformation reduces the linear model into the canonical form mentioned in the Introduction: z and $\bar{z}$ are independent, z has $N_p(\xi, \sigma^2 I_p)$ distribution, and $\bar{z}$ has $N_{n-p}(0, \sigma^2 I_{n-p})$ distribution.

The mapping between ξ, whose range is R^p, and η, whose range is the p-dimensional subspace $\mathcal{M}(X) \subset R^n$, is one-to-one:

$$\xi = U'\eta, \qquad \eta = U\xi. \tag{2}$$

Indeed, $\mathcal{M}(U) = \mathcal{M}(X)$ if and only if $X = UC$ for some $p \times p$ matrix C of rank p. From this and (1), $\xi = U'X\beta = U'UC\beta = C\beta$. Consequently, $\eta = X\beta = UC\beta = U\xi$.

Among the continuum of possible orthonormal bases for the regression space $\mathcal{M}(X)$, how should U be chosen? Computer packages for linear algebra offer numerically stable candidates that include: constructing U through the singular value decomposition $X = ULV'$; or taking $U = Q$ in the QR decomposition $X = QR$, where R is upper triangular; or using a standard orthonormal basis, such as the discrete Fourier transform in the special case $X = I_n$. In fact, nested principal component analysis relies on the singular value decomposition choice of U while nested order selection in polynomial regression may use the QR choice. See Section 2.4 for details.

Theoretical analysis in Section 4.3 indicates that the risk of a REACT fit is smaller if all but the first few components of ξ are very nearly zero. In this case, we say that the

orthogonal basis for the regression space is *economical* and will designate its matrix by U_E. Heuristically, the benefit of using an economical basis is clear. In that case, one need only identify and estimate the relatively few nonzero components of ξ, accumulating small squared biases from ignoring the nearly zero components but not accumulating the many variances that would arise from an attempt to estimate these unbiasedly. A basic flaw in the least squares fit is that it estimates unbiasedly every component of ξ, even those whose values are negligible.

The ideal choice of U_E would have its first column proportional to the unknown mean vector η, so that only the first component of ξ would be nonzero. Though unrealizable, this ideal choice makes the point that prior information or conjecture about the nature of η should be used in devising an economical basis matrix U_E for $\mathcal{M}(X)$. In many cases, it is likely that η is varies slowly between most pairs of adjacent components; and then it is plausible that the successive columns of an economical basis matrix are of increasing variation or, equivalently, of decreasing smoothness.

One-way layout. To develop this idea for the case of one-way layouts, let $D = \{d_{i,j}\}$ denote the first difference operator, the $(n-1) \times n$ matrix with $d_{i,i} = -1$, $d_{i,i+1} = 1$ and zeros elsewhere. Define the *roughness* of any vector $x \in R^n$ to be

$$V(x) = \sum_{i=2}^{n} (x_i - x_{i-1})^2 = |Dx|^2 \tag{3}$$

Slow variation in successive coordinates of η entails that $V(\eta)$ is small. Construct a decreasingly smooth basis for the regression space as follows:

a) Find an initial basis matrix U_0 for $\mathcal{M}(X)$ that has orthonormal columns. Numerically stable algorithms for the singular value decomposition or the QR decomposition provide convenient methods for this step.

b) Find the smoothest unit vector in $\mathcal{M}(X)$ by minimizing the roughness $V(U_0\gamma)$ over all $p \times 1$ unit vectors γ. This smoothest vector is evidently $U_0\gamma_p$, where γ_j denotes the eigenvector of $U_0'D'DU_0$ associated with the j-th largest eigenvalue λ_j.

c) Find the smoothest unit vector in $\mathcal{M}(X)$ that is orthogonal to the result of the previous step by minimizing $V(U_0\gamma)$ over all unit vectors γ that are orthogonal to γ_p. The answer is $U_0\gamma_{p-1}$.

d) Continue sequential constrained minimization to obtain the smooth basis matrix

$$U_S = (U_0\gamma_p, U_0\gamma_{p-1}, \ldots, U_0\gamma_1) = U_0\Gamma, \tag{4}$$

where $\Gamma = (\gamma_p, \gamma_{p-1}, \ldots, \gamma_1)$ is an orthogonal matrix. If $\Lambda = \mathrm{diag}\{\lambda_p, \lambda_{p-1}, \ldots, \lambda_1\}$, then

$$U_S'D'DU_S = \Gamma'U_0'D'DU_0\Gamma = \Lambda \tag{5}$$

and $U_S'U_S = I_p$. Equation (5) entails that the roughness of the k-th column of basis matrix U_S is equal to λ_{p-k+1}.

Examples discussed in Section 2.5 illustrate that the basis U_S is economical for many data sets taken from the literature on nonparametric smoothing when these are modelled as a one-way layout. Section 3.3 develops a smooth basis U_{SS} that is often economical for fitting a response surface to a two-way layout. This is not to say that all mean vectors η encountered in practice are smooth in the sense that $V(\eta)$ is small. Section 3.4 illustrates the differing economy of U_S in representing various signal types. In some cases, other bases, such as those related to wavelets, may be more economical. The REACT methodology can be expected to work effectively with any economical basis U_E.

Of particular interest is the trend model where $X = I_n$. Then the columns of the sequentially smooth basis matrix U_S described above are the eigenvectors of $D'D$, taken in increasing order of the associated eigenvalues. Moreover, as n increases, these eigenvectors converge swiftly to the *discrete cosine basis*, whose elements are the column vectors

$$\begin{aligned} c_1 &= \{n^{-1/2}: 1 \le j \le n\} \\ c_k &= \{(2/n)^{1/2} \cos[(2j-1)(k-1)\pi/(2n)]: 1 \le j \le n\} \quad \text{for } 2 \le k \le n. \end{aligned} \tag{6}$$

This analytical approximation works very well because the eigenvector property, that $D'Dc_k$ be proportional to c_k, holds exactly apart from the first and last elements of the vector $D'Dc_k$. In the context of Fourier analysis, the discrete cosine transform is a modification of the discrete Fourier transform that avoids creating Gibbs phenomena at the beginning and end of the REACT estimator of η. Rao and Yip (1990) discussed properties, algorithms and applications of the discrete cosine transform to digital signal processing.

2.2 Candidate estimators and estimated risks

Let U_E denote an economical basis for $\mathcal{M}(X)$. The one-to-one correspondence between the canonical mean $\xi = U_E'\eta$ and the original mean $\eta = U_E\xi$ carries over to estimators of these parameters. The risks of the paired estimators $\hat{\xi} = U_E'\hat{\eta}$ and $\hat{\eta} = U_E\hat{\xi}$ are identical:

$$R(\hat{\eta}, \eta, \sigma^2) = p^{-1}\mathrm{E}|\hat{\eta} - \eta|^2 = p^{-1}\mathrm{E}|\hat{\xi} - \xi|^2 = R(\hat{\xi}, \xi, \sigma^2). \tag{7}$$

In the canonical model, consider the linear estimators $\{\hat{\xi}(f) = fz: f \in \mathcal{F}\}$, where $\mathcal{F}$ is a specified subset of $[0,1]^p$. Such *candidate estimators* for ξ are also called modulation estimators or shrinkage estimators. The development here and in Section 2.3 draws on Beran and Dümbgen (1999) and Beran(1996).

For any $p \times 1$ vector h, let $\mathrm{ave}(h) = p^{-1}\sum_{i=1}^p h_i$. The risk of $\hat{\xi}(f)$ is

$$R(\hat{\xi}(f), \xi, \sigma^2) = \mathrm{ave}[\sigma^2 f^2 + \xi^2(1-f)^2] \equiv \rho(f, \xi^2, \sigma^2). \tag{8}$$

Define $\tilde{g} = \xi^2/(\xi^2 + \sigma^2)$, the operations being performed coordinatewise. Then $\tilde{g} \in [0,1]^p$ and

$$\rho(f, \xi^2, \sigma^2) = \mathrm{ave}[(f - \tilde{g})^2(\xi^2 + \sigma^2)] + \mathrm{ave}(\sigma^2\tilde{g}^2). \tag{9}$$

Ideally, if we knew the risk function in (8) and (9), we would use the candidate estimator $\hat{\xi}(\tilde{f}) = \tilde{f}z$, where

$$\tilde{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, \rho(f, \xi^2, \sigma^2) = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, \operatorname{ave}[(f - \tilde{g})^2(\xi^2 + \sigma^2)]. \tag{10}$$

Enlarging $\mathcal{F}$ to be a subset of R^p rather than $[0,1]^p$ does not reduce the minimal risk. Equation (9) shows that all minimizers of the risk are necessarily in $[0,1]^p$. When $\mathcal{F}$ is a convex subset of $[0,1]$, then $\tilde{f}$ is unique. Of special interest for the developments in this paper are three choices of $\mathcal{F}$:

The *global* class $\mathcal{F}_G = [0,1]^p$ is the largest possible. The value $\tilde{f}_G \in \mathcal{F}_G$ that minimizes risk is just $\tilde{g}$, defined above. The global class yields the ideal linear estimator $\tilde{\xi}_G = \tilde{g}z$ in the canonical model and $\tilde{\eta}_G = U_E \operatorname{diag}(\tilde{g}) U_E' y$ in the original parametrization.

The *monotone* class $\mathcal{F}_M$ is the convex set $\{f \in [0,1]^p : f_1 \geq f_2 \geq \ldots \geq f_p\}$. The importance of this class will become clearer in Section 2.3 on adaptation and in the asymptotic theory of Section 4. Note that it makes sense to shrink more severely the higher order components of z because the basis U_E provides an economical representation of η. The value $\tilde{f}_M \in \mathcal{F}_M$ that minimizes risk yields the ideal estimator $\tilde{\xi}_M = \tilde{f}_M z$ in the canonical model, which maps into $\tilde{\eta} = U_E \operatorname{diag}(\tilde{f}_M) U_E' y$ in the original parametrization.

To compute $\tilde{f}_M$, we use the right side of (10). If $\mathcal{H} = \{h \in R^p : : h_1 \geq h_1 \geq \ldots \geq h_p\}$, then

$$\tilde{f}_M = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \operatorname{ave}[(h - \tilde{g})^2(\xi^2 + \sigma^2)]. \tag{11}$$

This minimization is a weighted least squares isotonic regression problem, which may be solved numerically by the pooled adjacent violators (PAV) algorithm. For details of this algorithm, see Robertson, Wright and Dykstra (1988).

The *nested selection* class $\mathcal{F}_{NS}$ is the subset of $\mathcal{F}_{\mathcal{M}}$ defined as follows. For $0 \leq k \leq p$, let $e(k)$ denote the $p \times 1$ vector whose i-th component is 1 if $1 \leq i \leq k$ and is 0 otherwise. Then $\mathcal{F}_{NS} = \bigcup_{k=0}^{p} \{e(k)\}$. Because $\mathcal{F}_{NS}$ is a finite set that contains $p+1$ candidate values f, the computation of $\tilde{f}_{NS}$ is straightforward. In case of ties, we use as $\tilde{f}_{NS}$ the minimizing value of f that has the smallest number of nonzero entries. The resulting ideal estimator is $\tilde{\xi}_{NS} = \tilde{f}_{NS} z$ in the canonical model and $\tilde{\eta}_{NS} = U_E \operatorname{diag}(\tilde{f}_{NS}) U_E' y$ in the original parametrization. Sections 2.4 and 2.5 compare REACT fits with classical nested model selection, ridge, and principal component selection, none of which pay attention to economy of the regression basis in representing η.

2.3 Adaptation

The ideal linear estimators $\tilde{\xi}_G$, $\tilde{\xi}_M$, or $\tilde{\xi}_{NS}$ are usually unrealizable because ξ^2 and σ^2, which enter into the risk function $\rho(f, \xi^2, \sigma^2)$, are unknown. We therefore turn to the question of estimating risk. Three estimators of σ^2 prove useful for this purpose:

The *least squares* variance estimator is the classical residual-based estimator from least squares theory,

$$\hat{\sigma}^2_{LS} = (n-p)^{-1}|y - \hat{\eta}_{LS}|^2 = (n-p)^{-1}|\bar{z}|^2, \tag{12}$$

which is consistent if $n-p$ tends to infinity. The two following biased estimators are useful even when $p = n$. Their success relies on the economy of the basis U_E.

The *high component* variance estimator is

$$\hat{\sigma}^2_H = (n-n')^{-1}[\sum_{i=n'}^{p} z_i^2 + |\bar{z}|^2], \tag{13}$$

where $n' < p \leq n$. Because $\mathrm{E}\hat{\sigma}^2_H = \sigma^2 + (n-n')^{-1}\sum_{i=n'}^{p}\xi_i^2$, this estimator of σ^2 is consistent provided $n-n'$ tends to infinity and the bias term tends to zero. Economy of the basis U_E ensures that this bias term is relatively small.

The *first difference* variance estimator, treated by Rice (1984), is

$$\hat{\sigma}^2_D = [2(n-1)]^{-1}|Dy|^2 = [2(n-1)]^{-1}\sum_{i=2}^{n}(y_i - y_{i-1})^2. \tag{14}$$

In view of (5), the bias of this estimator for σ^2 is

$$[2(n-1)]^{-1}|D(\eta_i)|^2 = [2(n-1)]^{-1}\xi' U_S' D' D U_S \xi = [2(n-1)]^{-1}\xi'\Lambda\xi. \tag{15}$$

Consistency of $\hat{\sigma}^2_D$ is assured when n tends to infinity and the bias tends to zero. Because the basis U_E is economical, the smallness of the first few diagonal elements of Λ combines with the smallness of all but the first few components of ξ to control the right side of (15).

Having devised a consistent estimator $\hat{\sigma}^2$ of σ^2, we estimate ξ^2 by $z^2 - \hat{\sigma}^2$ and $\rho(f, \xi^2, \sigma^2)$ by

$$\hat{\rho}(f) = \mathrm{ave}[\hat{\sigma}^2 f^2 + (z^2 - \hat{\sigma}^2)(1-f)^2]. \tag{16}$$

The rationale for $\hat{\rho}(f)$ includes the calculation $\mathrm{E}z^2 = \xi^2 + \sigma^2$ and the supposition that the law of large numbers will make $\mathrm{ave}[(1-f)^2(z^2 - \hat{\sigma}^2)]$ consistent for $\mathrm{ave}[(1-f)^2\xi^2]$. Section 4.2 makes this precise. Only the manner in which σ^2 is estimated distinguishes the risk estimator $\hat{\rho}(f)$ from Stein's (1981) unbiased estimator of risk or from the Mallows (1973) C_L criterion. Define $\hat{g} = (z^2 - \sigma^2)/z^2$. Then $\hat{g} \in [-\infty, 1]^p$, not necessarily in $[0,1]^p$, and

$$\hat{\rho}(f) = \mathrm{ave}[(f - \hat{g})^2 z^2] + \mathrm{ave}(\hat{\sigma}^2 \hat{g}^2). \tag{17}$$

Adaptive estimation consists in using $\hat{\rho}(f)$ as a surrogate for the risk $\rho(f, \xi^2, \sigma^2)$ in identifying the best candidate estimator. Thus, for a given class $\mathcal{F}$ of shrinkage factors we consider the fully data-based estimator $\hat{\xi}(\hat{f}) = \hat{f}z$, where

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{\rho}(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \mathrm{ave}[(f - \hat{g})^2 z^2]. \tag{18}$$

We will write $\hat{\xi}_G$, $\hat{\xi}_M$, or $\hat{\xi}_{NS}$ to distinguish among the estimators of ξ that result from this construction when $\mathcal{F}$ is $\mathcal{F}_G$, $\mathcal{F}_M$, or $\mathcal{F}_{NS}$ respectively. The success or failure of this adaptation idea depends on the richness of the class $\mathcal{F}$. When $\mathcal{F} = \mathcal{F}_G$, it follows from (17) that the corresponding adaptive *global estimator* is $\hat{\xi}_G = \hat{g}_+ z$, where $\hat{g}_+$ is the positive part of $\hat{g}$. Unfortunately, the risk of $\hat{\xi}_G$ can be poor, unlike the risk of $\tilde{\xi}_G$. Suppose that $\hat{\sigma}^2$ and z are independent, as occurs when $\hat{\sigma}^2 = \hat{\sigma}^2_{LS}$. Under quadratic loss, the unique admissibility of z_i as an estimator of ξ_i entails that

$$\mathrm{E}[|\hat{\xi}_G - \xi|^2 | \hat{\sigma}^2] = \sum_{i=1}^{p} \mathrm{E}[\{(z_i^2 - \hat{\sigma}^2)_+ / z_i^2 - \xi_i\}^2 | \hat{\sigma}^2] > \sum_{i=1}^{p} \mathrm{E}[(z_i - \xi_i)^2 | \hat{\sigma}^2] = \sigma^2, \qquad (19)$$

for at least one value of ξ. At this ξ, the risk of $\hat{\xi}_G$ exceeds the risk of $\hat{\xi}_{LS}$. Thus, the risk function of $\hat{\xi}_G$ does not converge asymptotically to the risk function of $\tilde{\xi}_G$.

Adaptation works admirably for the smaller classes $\mathcal{F}_M$ and $\mathcal{F}_{NS}$, which yield, in the canonical parametrization, the adaptive *monotone estimator* $\hat{\xi}_M = \hat{f}_M z$ and the adaptive *nested selection estimator* $\hat{\xi}_{NS} = \hat{f}_{NS} z$. Section 4.2 describes how $\hat{\xi}_M$ and $\hat{\xi}_{NS}$ converge, both as estimators and in risk, to the ideal $\tilde{\xi}_M$ and $\tilde{\xi}_{NS}$ as p tends to infinity and $\hat{\sigma}^2$ converges in probability to σ^2. Experiments with artificial data suggest that the convergence of the adaptive estimators to their ideal counterparts is relatively quick.

Computing $\hat{\xi}_M$ is slightly more involved than computing the ideal $\tilde{\xi}_M$ in that

$$\hat{f}_M = \breve{f}_+ \text{ with } \breve{f} = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{ave}[(h - \hat{g})^2 z^2]. \qquad (20)$$

The positive-part step arises because $\hat{g}$ need not lie in $[0,1]^p$. For a proof of (20) as a consequence of (18), see Beran and Dümbgen (1999). The PAV algorithm provides an effective method for obtaining $\breve{f}$ and hence $\hat{f}_M$. Computing $\hat{f}_{NS}$ is straightforward minimization over a finite set. In the original parametrization, the two adaptive estimators become $\hat{\eta}_M = U_E \mathrm{diag}(\hat{f}_M) U_E' y$ and $\hat{\eta}_{NS} = U_E \mathrm{diag}(\hat{f}_{NS}) U_E' y$.

2.4 Connections

The adaptive estimators in Section 2.3 are defined to minimize estimated risk, or equivalently, a C_L criterion. Mallows (1973) noted heuristically that the size of $\mathcal{F}$ affects the success or failure of minimum C_L. Li and Hwang (1984) presented Stein-type shrinkage estimators that dominate $\hat{\eta}_{LS}$ in risk. Li (1985) established for nested model selection, ridge regression, and certain other examples the convergence of $\hat{\rho}(f)$ (with σ^2 assumed known) to the loss of $\hat{\xi}(f)$, uniformly over f in $\mathcal{F}$. Kneip (1994) gave related results for the larger class of ordered linear smoothers. The asymptotic equivalence in loss of estimators obtained by minimizing Stein's unbiased estimator of risk, or the generalized cross-validation criterion, or the C_L criterion was explored by Li (1985, 1987).

On the other hand, Efroimovich and Pinsker (1984) and Golubev (1987) constructed adaptive estimators whose maximum risk converges asymptotically to the Pinsker (1980)

bound for each member of a class of ellipsoids in the parameter space. Beran and Dümbgen (1999) developed conditions on the covering number of $\mathcal{F}$ that ensure correct convergence of the loss and risk of $\hat{\eta}_{\mathcal{F}}$ to their counterparts for $\tilde{\eta}_{\mathcal{F}}$ and linked their results to Pinsker theory. The Pinsker asymptotic minimaxity of $\hat{\xi}_M$ and $\hat{\xi}_{NS}$ are discussed further in Section 4. This section compares REACT fits with several precursors and competitors.

Nested polynomial regression. Suppose that the columns of the regression matrix X are the powers $\{x^k: 0 \le k \le p-1\}$ of an $n \times 1$ covariate vector x. The QR decomposition of the regression matrix is $X = QR$, where the columns of the $n \times p$ matrix Q are orthonormal and R is upper triangular. Taking the orthonormal basis for the regression space to be Q, reduce the linear model to canonical form as in Section 2.1 and consider the fit $Q\mathrm{diag}(\hat{f}_{NS})Q'y$, which is the adaptive nested selection estimator in the canonical model, mapped back into the original parametrization. Because the QR decomposition expresses the Gram-Schmidt orthogonalization of X, this adaptive estimator is equivalent to choosing the order of the original polynomial regression by minimizing the C_L criterion and then fitting the polynomial of this order by least squares.

To this extent, polynomial regression with order chosen to minimize the C_L criterion is a precursor to REACT. Moreover, such an adaptive polynomial fit has smaller asymptotic risk than the least squares estimator $\hat{\eta}_{LS}$ because the asymptotic theory in Section 4 applies to any canonical form of the linear model. However, because the basis matrix Q obtained from the QR decomposition of polynomial X need not be economical for smooth signals, the reduction in risk may be small. This difficulty is illustrated by the unsuccessful polynomial fits to the motorcycle data displayed on p. 325 of Venables and Ripley (1997).

Nested principal component selection. The singular value decomposition of an $n \times p$ regression matrix X is $X = ULV'$, where U is $n \times p$, V is $p \times p$, $U'U = V'V = VV' = I_p$ and and $L = \mathrm{diag}\{l_i: 1 \le i \le p\}$ with $l_1 \ge l_2 \ge \ldots \ge l_p > 0$. The columns of V are eigenvectors of $X'X$. Rao and Toutenberg (1995), p. 62, formulated nested principal component selection as follows. The mean $X\beta$ of the linear model can be rewritten as $\tilde{X}\tilde{\beta}$, where $\tilde{X} = XV$ and $\tilde{\beta} = V'\beta$. Let $\tilde{X}_k$ denote the $n \times k$ matrix formed from the first k columns of $\tilde{X}$. Candidate nested principal components estimators for η are defined by

$$\hat{\eta}(k) = \tilde{X}_k(\tilde{X}_k\tilde{X}_k')^{-1}\tilde{X}_k y \tag{21}$$

for $1 \le k \le p$ and $\hat{\eta}(0) = 0$.

Applying the singular value decomposition to (21) yields the equivalent expression $\hat{\eta}(k) = U\mathrm{diag}(e(k))U'y$, where $e(k)$ is the vector of k ones and $n-k$ zeros defined in the paragraph that follows (11). Thus, if k is chosen to minimize the C_L criterion, nested principal component regression is analogous to the adaptive nested selection estimator of Section 2.3, with the principal component basis in place of U_E. Because the asymptotic theory in Section 4 applies to any canonical form of the linear model, the adaptive nested principal component

fit has no greater asymptotic risk than the least squares fit $\hat{\eta}_{LS}$. However, the uncertain success of principal component regression in applications stems from its use of an orthogonal basis that does not attempt an economical representation of η. Section 2.5 illustrates this difficulty in fitting one-way layouts to two well-known sets of data.

Ridge regression. In the notation of the singular value decomposition for X, the candidate estimators for η in ridge regression are

$$\hat{\eta}(c) = X(X'X + cI_p)^{-1}X'y = Uf(c)U'y, \tag{22}$$

where $c \geq 0$ and $f(c) = \{l_i^2/(l_i^2 + c): 1 \leq i \leq p\}$. Evidently the range of the candidate shrinkage vectors $\{f(c): c \geq 0\}$ is a proper subset of $\mathcal{F}_M$. Thus, if c is chosen to minimize the C_L criterion, the resulting ridge regression estimator $\hat{\eta}_{RIDGE}$ has no greater asymptotic risk than $\hat{\eta}_{LS}$ (see Section 4). However, because it tacitly uses the principal component basis without regard to the economy of that basis in representing η, ridge regression may not improve significantly upon least squares. See Section 2.5 for examples.

Symmetric linear smoothers. In REACT, the candidate estimators for η take the form $\hat{\eta} = U_E \text{diag}(f)U_E'y$, where $f \in \mathcal{F} \subset [0,1]^p$ and the economical regression basis U_E depends upon X. The matrix $A = U_E \text{diag(f)} U_E'$ is symmetric with eigenvalues restricted to $[0,1]$ and does not depend on y. The candidate estimators are thus symmetric linear smoothers, in the terminology of Buja, Hastie, and Tibshriani (1989). For given linear smoother, that paper identifies the matrix A and analyzes its singular value decomposition, which reduces to the spectral decomposition when A is symmetric.

The REACT approach is synthetic, defining a class of candidate symmetric matrices A through a class of possible eigenvalues $f \in \mathcal{F}$ and through the eigenvectors U_E. The restriction to an economical basis U_E when specifying candidate values of A is motivated by efficiency arguments in Sections 2.1 and 4.3. The main thrusts of this paper are: (a) justifying theoretically the use of an economical basis U_E followed by adaptive selection of f through minimization of the estimated risk $\hat{\rho}(f)$; (b) showing empirically that the smooth basis U_S is often economical for one-way layouts; (c) developing and probing confidence sets for η that are centered at REACT fits. The asymptotics in Section 4 also support adaptation over a finite collection of plausible economical bases. Unlike the candidate estimators, the REACT estimator $\hat{\eta}_{\mathcal{F}} = U_E \text{diag}(\hat{f}_{\mathcal{F}})U_E'y$ is nonlinear in y because $\hat{f}_{\mathcal{F}}$ depends on both y and X. Mallows (1980) treated properties of nonlinear smoothers in a random signal model. His framework contrasts with the present analysis of REACT estimators under a linear model having deterministic regressors.

Sparse bases and hard thresholding. A sparse basis is one in which only a few basis vectors are needed to obtain a good approximation to η. An economical basis, as described in Section 2.1, is a sparse basis in which the first few basis vectors provide the good approximation. Donoho and Johnstone (1994) studied hard-thresholding estimators of the form $\hat{\xi}_i = z_i$ if

$|z_i| \geq \lambda_p \sigma^2$ and $\hat{\xi}_i = 0$ otherwise. Such nonlinear shrinkage estimators were shown to have surprisingly small risk when the canonical model is sparse and $\lambda_p/[2\log(p)]^{1/2}$ tends to 1 as p increases. A numerical experiment in Beran and Dümgben (1999) suggested that the success of hard-thresholding estimators may be more sensitive to the choice of basis than the success of adaptive linear shrinkage estimators. Confidence sets based on hard-thresholding estimators appear to be an open question.

2.5 REACT on Data

After the initial choices of regression matrix, basis, variance estimator, and shrinkage class $\mathcal{F}$, REACT fits are completely defined through the process of minimizing estimated risk. There is no need to guess or estimate bandwidth parameters. This and further points will now be illustrated through examples.

Motorcycle data. Competing nonparametric smoothing fits to the motorcycle data are displayed on p. 325 of Venables and Ripley (1997), on p. 97 of Fan and Müller (1995), and on pp. 8–11 of Silverman (1985). Conditioning on the observed times, we will fit an unbalanced one-way layout to the $n = 133$ observed accelerations, the factor levels being the $p = 94$ times taken in increasing order. Each row in the regression matrix X thus contains a single 1, the other elements all being 0. Repeated observations at a time point induce repeated rows in X. Because of replication, it is reasonable to estimate σ^2 by the least squares estimator $\hat{\sigma}^2_{LS} = 599.5$ defined in (12).

[Figures 1, 2 and 3 go near here]

The right column in Fig. 1 displays the least squares, adaptive nested principal component, and adaptive ridge fits to this linear model, line segments being drawn between the successive fitted means. Minimum estimated risk (or equivalently, minimum C_L) was used to select the ridge parameter and the number of principal components used. Visually, the latter two fits are no improvement over the unconvincing least squares fit. Clearly evident is the tendency of least squares to overfit whenever p is not small.

The left column in Fig. 1 exhibits the motorcycle data and two REACT fits that use the smooth basis U_S: the adaptive nested selection fit $\hat{\eta}_{NS}$ and the adaptive monotone fit $\hat{\eta}_M$, both defined in Section 2.3. Line segments are drawn between the means fitted in this manner to the one-way layout. It is striking how well these two REACT fits to the motorcycle data compare visually with the best competing fits in the literature cited above.

The estimated risks for the various fits in Fig. 1 reveal the heart of the story. On the one hand, $\hat{\rho}_{LS} = 599.5$, $\hat{\rho}_{PC} = 423.2$, and $\hat{\rho}_{RIDGE} = 497.0$, all of which are similarly high. In sharp contrast the two REACT fits have $\hat{\rho}_{NS} = -75.2$ and $\hat{\rho}_M = -76.4$. The negative values cause no concern because the risk estimator, defined in (16), is not constrained to be positive. Of interest are three points: (a) both REACT fits have much smaller risk than the least squares, nested principal component, or ridge fits; (b) in terms of estimated risk, the

nested selection smooth fit does nearly as well as the monotone smooth fit in this example; (c) much smaller estimated risk corresponds to a better visual fit.

The first row of Fig. 2 presents two diagnostic plots: on the left, the canonical response $z = U_S'y$ and, on the right, the adaptive shrinkage factors $\hat{f}_{NS}$ and $\hat{f}_M$. The relatively small magnitude of all but the first few components of z supports the belief that the smooth basis U_S represents η economically. Note that the plot of z displays the square roots $\{z_i^{1/2}\}$, so as to better scrutinize values near 0. The close similarity of the two $\hat{f}$ plots explains the near coincidence of the two smooth REACT fits in Fig. 1. As might be expected from the discussion in Section 2.1, the first four basis vectors in U_S, displayed in the second and third rows of Fig. 2, are a perturbation of the first four vectors in the discrete cosine basis. The flat steps in the basis vectors reflect repeated observations at some of the factor levels of the one-way layout being fitted to the motorcycle data.

Fig. 3 presents analogous diagnostic plots for the principal component basis that underlies the principal component (PC) and ridge fits in Fig. 1. It is clear from the $(1,1)$ plot that that the PC basis does not represent the mean acceleration economically. Consequently, as displayed in the (1,2) plot, most basis vectors are retained when minimizing estimate risk. This circumstance precludes much reduction in risk by either nested principal component or ridge regression. On looking at the first four vectors in the principal component basis, this lack of success is not entirely surprising.

Redoing the analysis of the motorcycle data with the first difference variance estimator $\hat{\sigma}_D^2$ in place of $\hat{\sigma}_{LS}^2$ makes no perceptible difference to the plotted fits.

Geyser data. Simonoff (1996), pp. 135–6, smoothed nonparametrically the Old Faithful geyser data. Conditioning on the observed eruption durations, we will fit an unbalanced one-way layout to the $n = 222$ observed eruption intervals, the factor levels being the $p = 34$ eruption durations taken in increasing order. The regression matrix X is analogous to the one used for the motorcycle data and the variance σ^2 is reasonably estimated by $\hat{\sigma}_{LS}^2 = 35.6$.

[Figures 4 and 5 go near here]

Figs. 4 and 5 are counterparts for the geyser data of Figs. 1 and 2 for the motorcycle data. Visually, the least squares, principal component and ridge fits in Fig. 4 are virtually identical and are far less satisfactory than the nested selection and monotone fits that use the smooth basis U_S. Of the two REACT fits, the monotone selection fit seems slightly better in capturing nuances of the data. The estimated risks for the various fits agree with the visual impressions. On the one hand, $\hat{\rho}_{LS} = 35.6$, $\hat{\rho}_{PC} = 35.6$, and $\hat{\rho}_{RIDGE} = 35.5$. On the other hand, $\hat{\rho}_{NS} = 5.2$ and $\hat{\rho}_M = 4.5$.

The diagnostic plot of z in Fig. 5 supports the belief that the smooth basis U_S represents η economically. The $\hat{f}$ plots show that the monotone smooth fit reduces risk over the nested selection fit by using additional, greatly shrunk, components of z. As might be expected, the first four basis vectors in U_S, displayed in the second and third rows of Fig. 2, are

a perturbation of the first four vectors in the discrete cosine basis. Their form reflects repeated observations at many factor levels of the one-way layout.

Redoing the analysis of the geyser data with the first difference variance estimator $\hat{\sigma}_D^2$ in place of $\hat{\sigma}_{LS}^2$ makes no visible difference to the plotted fits.

3. EXTENSIONS

The discussion in Section 2 focused on REACT fits to the one-way layout with homoscedastic errors. This section fits the two-way layout, deals with a simple form of heteroscedasticity, constructs and explores confidence sets for η, and looks numerically at the role of an economical basis in REACT.

3.1. Two-way layout

The coal ash data from p. 34 of Cressie (1993) will be fitted as an incomplete two-way layout with $n = p = 208$, the factor pairs being the grid coordinates at which the measurements are taken. To obtain an economical basis, the concept of smoothness that was used in Section 2 for the one-way layout now needs to be extended. Let C denote the subset of factor level pairs for which there is a measurement y_{ij}. For any incomplete matrix $A = \{a_{ij}: (i,j) \in C\}$, let $D_{i\cdot}A$ denote the vector of first differences computed from the i-th row of A, ignoring missing elements. Similarly, let $D_{\cdot j}A$ denote the vector of first differences computed from the j-th column of A. The roughness of the mean matrix $\eta = \{\eta_{ij}: (i,j) \in C\}$ is now defined to be

$$V(\eta) = \sum_{\text{all } i} |D_{i\cdot}\eta|^2 + \sum_{\text{all } j} |D_{\cdot j}\eta|^2. \tag{23}$$

If we systematically rearrange the matrix η, row by row, into a long vector $\text{vec}(\eta)$, then $V(\eta) = |D\text{vec}(\eta)|^2$, where D is a matrix each of whose rows contains a single 1 and a single -1, the other row entries all being 0. With this change in the definition of D, we now construct a smooth basis U_{SS} for the regression space of the two-way layout by the method described in Section 2.1. Fig. 7 displays the first six members of U_{SS} for the coal ash data as surfaces on C, with linear interpolation between grid points. When the two-way layout is complete, the basis U_{SS} is asymptotically equivalent, as both dimensions of the two-way layout increase, to a bivariate discrete cosine transform. The basis vectors in Fig. 7 are visibly related to this transform.

[Figures 6 and 7 go near here]

Let $y = \{y_{ij}: (i,j) \in C\}$ and let n denote the cardinality of C. Adapted to the two-way layout, the first difference variance estimator becomes

$$\hat{\sigma}_D^2 = [2(n-1)]^{-1} |D\text{uvec}(y)|^2, \tag{24}$$

where D is now defined as in the preceding paragraph, not as in (14). The least squares

estimator of σ^2 is not available for the coal ash data because there is only one observation per cell.

The left column in Fig. 6 exhibits the least squares fit to the coal ash data (i.e. the data itself with linear interpolation between grid points) and two REACT fits that use the smooth basis U_{SS}: the adaptive nested selection fit $\hat{\eta}_{NS}$ and the adaptive monotone fit $\hat{\eta}_M$, both defined as in Section 2.3 after setting $U_E = U_{SS}$. The estimated risks for these fits are $\hat{\rho}_{LS} = 1.03$, $\hat{\rho}_{NS} = .17$ and $\hat{\rho}_M = .12$. Not only are the two REACT fits more pleasing visually; they also have much smaller estimated risk than the least squares fit.

The diagnostic plot of z in the right column of Fig. 6 indicates that the smooth basis U_{SS} represents η economically. The $\hat{f}$ plots help explain why the monotone and nested selection smooth fits to the coal ash data are similar, although the monotone fit uses additional, greatly shrunk, components of z to achieve higher fidelity to the data without overfitting.

3.2. Heteroscedasticity

The first 21 accelerations in the motorcycle data of Fig. 1 appear to have much smaller variability than the other accelerations. This is a consequence of the experimental conditions, some details of which are reported by Silverman (1985). To take the possible change in variability into account, we divide the sample into two subsamples, consisting of the first 21 and the remaining 112 data points. The aim is to fit a separate one-way layout to each subsample by the REACT method, using the appropriate smooth basis U_S and monotone shrinkage for each subsample. The first difference variance estimator for subsample 1 and the least squares variance estimator for subsample 2 provide the estimated variances required for this procedure. The right plot in Fig. 8 is the result after linear interpolation between adjacent fitted points. The left plot in Fig. 8 is the monotone smooth basis fit obtained on the assumption that the data is homoscedastic (see also Fig. 1). The fit under the heteroscedastic model better captures the abrupt change in acceleration from zero to negative.

[Figure 8 goes near here]

3.3. Probing confidence sets

To construct a confidence set for the unknown mean vector η that is centered at the estimator $\hat{\eta}_{\mathcal{F}}$, consider the root

$$\hat{t}_{\mathcal{F}} = p^{1/2}[p^{-1}|\hat{\eta}_{\mathcal{F}} - \eta|^2 - \hat{\rho}(\hat{f}_{\mathcal{F}})], \tag{25}$$

where $\mathcal{F}$ is either the monotone class $\mathcal{F}_M$ or the nested selection class $\mathcal{F}_{NS}$. The right side of (25) compares the normalized quadratic loss $L(\hat{\eta}_{\mathcal{F}}, \eta) = L(\hat{\xi}_{\mathcal{F}}, \xi)$ with the estimated risk of $\hat{\eta}_{\mathcal{F}}$ or $\hat{\xi}_{\mathcal{F}}$. As discussed in Section 4, the loss and risk of $\hat{\xi}_{\mathcal{F}}$ converge together when p increases. A confidence set for $\hat{\eta}_{\mathcal{F}}$ is obtained by referring $\hat{t}_{\mathcal{F}}$ to the α-th quantile of its estimated distribution. The general idea behind such confidence sets was sketched in the last paragraphs of Stein (1981).

Further details of the construction depend on the the variance estimator that enters into the definition of $\hat{t}_{\mathcal{F}}$. We take $\hat{\sigma}^2 = \hat{\sigma}^2_{LS}$ in (25). For large p and $n-p$, the distribution of $\hat{t}_{\mathcal{F}}$ is then approximately $N(0, \hat{\tau}^2_{\mathcal{F}})$ with

$$\begin{aligned}\hat{\tau}^2_{\mathcal{F}} = 2\hat{\sigma}^4_{LS}\text{ave}[(2\hat{f}_{\mathcal{F}} - 1)^2] + 2[p/(n-p)]\hat{\sigma}^4_{LS}[\text{ave}(2\hat{f}_{\mathcal{F}} - 1)]^2 \\ + 4\hat{\sigma}^2_{LS}\text{ave}[(z^2 - \hat{\sigma}^2_{LS})(1 - \hat{f}_{\mathcal{F}})^2].\end{aligned} \tag{26}$$

A confidence set of approximate coverage probability α for η is accordingly

$$\hat{C}_{\mathcal{F}} = \{\theta \in \mathcal{M}(X): |\hat{\eta}_{\mathcal{F}} - \theta|^2 \le p\hat{\rho}(\hat{f}_{\mathcal{F}}) + p^{1/2}\hat{\tau}_{\mathcal{F}}\Phi^{-1}(\alpha)\}. \tag{27}$$

Here, as usual, Φ^{-1} denotes the quantile function of the standard normal distribution. Section 4.4 presents the underlying asymptotic theory.

When the construction is carried out for the least squares estimator $\hat{\eta}_{LS}$, the root in (25) simplifies to

$$\hat{t}_{LS} = p^{1/2}[p^{-1}|\hat{\eta}_{LS} - \eta|^2 - \hat{\sigma}^2_{LS}]. \tag{28}$$

The approximate distribution of $\hat{t}_{LS}$ for large p and $n-p$ is now $N(0, \hat{\tau}^2_{LS})$, with $\hat{\tau}^2_{LS} = 2\hat{\sigma}^4(n/p)$. The confidence set for η centered at $\hat{\eta}_{LS}$ is thus

$$\hat{C}_{LS} = \{\theta \in \mathcal{M}: |\hat{\eta}_{LS} - \theta|^2 \le p\hat{\sigma}^2_{LS} + p^{1/2}(2n/p)^{1/2}\hat{\sigma}^2_{LS}\Phi^{-1}(\alpha)\}. \tag{29}$$

If p and $n-p$ both tend to infinity and $p/(n-p)$ converges to a finite constant, then $\hat{C}_{LS}$ approximates the classical confidence set that refers $(p\hat{\sigma}^2_{LS})^{-1}|\hat{\eta}_{LS} - \eta|^2$ to the α-th quantile of the F distribution with p and $n-p$ degrees of freedom.

When $\mathcal{F}$ is either the monotone or nested selection class and $\alpha \in (0,1)$ is fixed, the maximum distance between η and elements of the confidence set is asymptotically smaller for $\hat{C}_{\mathcal{F}}$ than for $\hat{C}_{LS}$ (see Section 4.4). Unfortunately, visualizing either of these confidence sets for η is difficult. In the canonical parametrization, $\hat{C}_{\mathcal{F}}$ simplifies to a ball in R^p centered at the estimated canonical mean $\hat{\xi}_{\mathcal{F}}$. However, the canonical confidence ball, like the original confidence ball, lacks convenient interpretation in the examples of Section 2.5.

A useful idea is to probe the extent of the confidence set by identifying extreme elements in $\hat{C}_{\mathcal{F}}$. For instance, we can refit the model after replacing $\hat{\sigma}^2_{LS}$ with $c\hat{\sigma}^2_{LS}$, where c is a positive constant. For monotone or nested selection fits using a smooth basis, the smoothness of the estimator $\hat{\eta}_{\mathcal{F}}(c)$ increases with c. Let

$$\hat{c}_L = \inf\{c > 0: \hat{\eta}_{\mathcal{F}}(c) \in \hat{C}_{\mathcal{F}}\}, \qquad \hat{c}_U = \sup\{c > 0: \hat{\eta}_{\mathcal{F}}(c) \in \hat{C}_{\mathcal{F}}\}. \tag{30}$$

Then $\hat{\eta}_{\mathcal{F}}(\hat{c}_L)$ and $\hat{\eta}_{\mathcal{F}}(\hat{c}_U)$ are, respectively, the roughest and smoothest rescaled-variance fits that lie in the confidence set.

[Figure 9 goes near here]

Fig. 9 displays (solid line) roughest and smoothest rescaled-variance refits for the motorcycle data previously analyzed. These two refits lie on the boundary of the asymptotic 90% confidence set for η that is centered at the monotone smooth basis fit $\hat{\eta}_M$ (dashed line) from Fig. 1. One can reasonably have confidence in the existence of broad features preserved by the two extreme refits.

3.4. Role of an economical basis

Both the heuristic considerations of Section 2.1 and the asymptotics in Section 4 indicate that REACT reduces risk most when the basis for the regression space is economical. To see directly the effects of economy on the fit, we consider three numerical examples based on artificial data. In each case the regression matrix is the identity I_n. The i-th component of the mean vector η takes the form $\eta_i = m(i/(n+1))$, $1 \le i \le n$, where m is a function defined on the interval $[0,1]$. Three choices for m are considered:

Smooth: $m_1(t) = 2(6.75x^2(1-x))^3$.

Burst: $m_2(t) = 0$ if $0 \le t \le .25$ and $= \sin(2\pi/t)$ if $.25 < t \le 1$.

Steps: $m_3(t) = 0$ if $0 \le t \le .15$, $= 1.5$ if $.15 < t \le .3$, $= .5$ if $.3 < t \le .6$, $= -.5$ if $.6 < t \le .8$, and $= 1$ if $.8 < t \le 1$.

[Figure 10 goes near here]

The left column of Fig. 10 displays for each j a pseudo-random sample of size $n = 200$ in which the i-th observation is drawn from the $N(m_j(i/(n+1)), \sigma^2)$ distribution with $\sigma = .2$. The dashed line plots in this column represent the respective vectors η, with linear interpolation between components. The solid line plots similarly display, for the discrete cosine basis U_{DC}, monotone fits to the three samples. Comparison of the fits with the data and with the true η brings out several points: (a) the REACT fits quickly track sharp changes in η; (b) the fits to the second and third cases are rougher than the fit to the first case; (c) wiggles in the fits accurately reflect patterns in the data. These findings are not very sensitive to the choices of n and σ^2 in the numerical experiment.

The right column in Fig. 10 displays the components of the canonical mean vector $\xi = U_{DC}'\eta$ for each of the three cases. These diagnostic plots show that the discrete cosine basis is substantially more economical for the first η than for the other two. Consequently, in the second and third cases, the monotone REACT estimator shrinks the higher-order components of z more conservatively than in the first case. This explains point (b) in the preceding paragraph.

Further numerical experimentation reveals that the monotone Fourier-basis fit to the step function data suffers from Gibbs phenomena jumps at its two endpoints. This difficulty arises because the endpoints of the true η are not equal. The discrete cosine basis avoids such end effects, as does the smooth basis U_S more generally for one-way layouts.

4. SUPPORTING ASYMPTOTICS

Several theoretical results on the performance of REACT fits are the topic of this section. The asymptotics require that the dimension p of the regression space tend to infinity and, at a minimum, that the variance estimator $\hat{\sigma}^2$ be consistent for σ^2. Further properties are required of the variance estimator in constructing confidence sets for η.

4.1. Minimax bounds

The analysis in Pinsker's (1980) paper yields the two minimax theorems stated below for estimation in a canonical linear model. The formulation is that of Section 2, the quadratic risk $R_p(\hat{\xi}, \xi, \sigma^2)$ being defined by (7). Both theorems hold for every possible selection of the orthonormal basis U in (1). How the choice of basis affects the minimax risk is studied in Section 4.3.

Let $\mathcal{E} = \{a \in R^p : a_i \in [1, \infty], 1 \le i \le p\}$. For every $a \in \mathcal{E}$, define the ellipsoid

$$E(r, a, \sigma^2) = \{\xi \in R^p : \text{ave}(a\xi^2) \le \sigma^2 \sigma^2\}. \tag{31}$$

If $\xi \in E(r, a, \sigma^2)$ and $a_i = \infty$, it is to be understood that $\xi_i = 0$ and $a_i^{-1} = 0$. Let

$$\xi_0^2 = \sigma^2[(\mu/a)^{1/2} - 1]_+ \qquad g_0 = \xi_0^2/(\sigma^2 + \xi_0^2) = [1 - (a/\mu)^{1/2}]_+, \tag{32}$$

where μ is the unique positive number such that $\text{ave}(a\xi_0^2) = \sigma^2 r$. Define

$$\nu_p(r, a, \sigma^2) = \rho(g_0, \xi_0^2, \sigma^2) = \sigma^2 \text{ave}[\xi_0^2/(\sigma^2 + \xi_0^2)]. \tag{33}$$

Evidently, $\nu_p(r, a, \sigma^2) \in [0, \sigma^2]$ for every $r > 0$ and every $a \in \mathcal{E}$.

The first theorem specialized from Pinsker's argument identifies the linear estimator that is minimax among all linear estimators of ξ and finds the minimax risk for this class.

Theorem 1. For every $a \in \mathcal{E}$ and every $r > 0$,

$$\inf_{f \in R^p} \sup_{\xi \in E(r,a,\sigma^2)} R_p(fz, \xi, \sigma^2) = \nu_p(r, a, \sigma^2) = \sup_{\xi \in E(r,a,\sigma^2)} R(g_0 z, \xi, \sigma^2). \tag{34}$$

The second theorem establishes that the minimax linear estimator $g_0 z$ is often asymptotically minimax among all estimators of ξ.

Theorem 2. For every $a \in \mathcal{E}$ and every $r > 0$ such that $\lim_{p\to\infty} p\nu_p(r, a, \sigma^2) = \infty$,

$$\lim_{p\to\infty} [\inf_{\hat{\xi}} \sup_{\xi \in E(r,a,\sigma^2)} R_p(\hat{\xi}, \xi, \sigma^2)/\nu_p(r, a, \sigma^2)] = 1 \tag{35}$$

and

$$\lim_{p\to\infty} [\sup_{\xi \in E(r,a,\sigma^2)} R_p(g_0 z, \xi, \sigma^2)/\nu_p(r, a, \sigma^2)] = 1. \tag{36}$$

If $\liminf_{p\to\infty} \nu_p(r, a, \sigma^2) > 0$, then also

$$\lim_{p\to\infty} [\inf_{\hat{\xi}} \sup_{\xi \in E(r,a,\sigma^2)} R_p(\hat{\xi}, \xi, \sigma^2) - \nu_p(r, a, \sigma^2)] = 0 \tag{37}$$

and

$$\lim_{p\to\infty}[\sup_{\xi\in E(r,a,\sigma^2)} R_p(g_0 z,\xi,\sigma^2) - \nu_p(r,a,\sigma^2)] = 0. \tag{38}$$

Because g_0 depends on r, a, and σ^2, the asymptotic minimaxity of $g_0 z$ is assured only over the one ellipsoid $E(r,a,\sigma^2)$. The following idealized estimator, which depends on ξ^2 and σ^2, is asymptotically minimax over a class of such ellipsoids. Let $\mathcal{E}_0 \subset \mathcal{E}$ and $\mathcal{F}$ be such that $g_0(r,a,\sigma^2) \in \mathcal{F}$ for every $a \in \mathcal{E}_0$, every $r > 0$, and every $\sigma^2 > 0$. For the sake of successful adaptation in the next subsection, we desire that the shrinkage class $\mathcal{F}$ be not too large. This requirement limits the choice of $\mathcal{E}_0$. Because both $\tilde{f}$ and g_0 lie in $\mathcal{F}$, it follows from (10) that

$$\sup_{\xi\in E(r,a,\sigma^2)} R_p(\tilde{f}z,\xi,\sigma^2) \le \sup_{\xi\in E(r,a,\sigma^2)} R_p(g_0 z,\xi,\sigma^2) \tag{39}$$

for every $a \in \mathcal{E}_0$, every $r > 0$ and every $\sigma^2 > 0$. Thus, if g_0 is replaced by $\tilde{f}$, the limits (36) and (38) continue to hold for every $a \in \mathcal{E}_0$, every $r > 0$ and every $\sigma^2 > 0$. This establishes the asymptotic minimaxity of $\tilde{f}z$ over the class of ellipsoids $E(r,a,\sigma^2)$ generated as a ranges over $\mathcal{E}_0$ and r ranges over the positive reals.

4.2. Adaptation

As described in Section 2.3, adaptation consists in using the estimator $\hat{f}z$, which depends only on the data, as a surrogate for the idealized estimator $\tilde{f}z$. Equation (18) defines $\hat{f}$. The following result, which specializes Theorems 2.1 and 2.2 in Beran and Dümbgen (1999), gives sufficient conditions on $\mathcal{F}$ and $\hat{\sigma}^2$ to ensure that $\hat{f}z$ behaves asymptotically like $\tilde{f}z$.

Theorem 3. Let $\mathcal{F}$ be any subset of $\mathcal{F}_M$ that is closed in $[0,1]^p$ and contains the vector 0. Suppose that $\hat{\sigma}^2$ is consistent in that, for every $r > 0$ and $\sigma^2 > 0$,

$$\lim_{p\to\infty} \sup_{\mathrm{ave}(\xi^2)\le\sigma^2 r} \mathrm{E}|\hat{\sigma}^2 - \sigma^2| = 0. \tag{40}$$

Then, for every $r > 0$ and every $\sigma^2 > 0$,

$$\lim_{p\to\infty} \sup_{\mathrm{ave}(\xi^2)\le\sigma^2 r} \mathrm{E}\sup_{f\in\mathcal{F}} |\hat{\rho}(f) - \rho(f,\xi^2,\sigma^2)| = 0. \tag{41}$$

Moreover, the estimators $\hat{\xi}_{\mathcal{F}} = \hat{f}z$ and $\tilde{\xi}_{\mathcal{F}} = \tilde{f}z$ satisfy, for every $r > 0$ and $\sigma^2 > 0$,

$$\lim_{p\to\infty} \sup_{\mathrm{ave}(\xi^2)\le\sigma^2 r} |R_p(\hat{\xi}_{\mathcal{F}},\xi,\sigma^2) - R_p(\tilde{\xi}_{\mathcal{F}},\xi,\sigma^2)| = 0 \tag{42}$$

and

$$\lim_{p\to\infty} \sup_{\mathrm{ave}(\xi^2)\le\sigma^2 r} \mathrm{E}\,\mathrm{ave}[(\hat{\xi}_{\mathcal{F}} - \tilde{\xi}_{\mathcal{F}})^2] = 0. \tag{43}$$

This theorem gives conditions under which the adaptive estimator $\hat{\xi}_{\mathcal{F}}$ and the idealized estimator $\tilde{\xi}_{\mathcal{F}}$ converge together as random vectors and in risk. The hypothesis on

the shrinkage class $\mathcal{F}$ includes both $\mathcal{F}_M$ and $\mathcal{F}_{NS}$. Condition (40) on the variance estimator holds for $\hat{\sigma}^2_{LS}$ if $n-p$ tends to infinity with p. The same condition holds for $\hat{\sigma}^2_D$ if $\lim_{p\to\infty} p^{-1}\sum_{i=2}^{p}(\eta_i-\eta_{i-1})^2 = 0$. This conclusion follows from (15); for details see Beran (1996).

4.3. Effect of basis choice

The results of Sections 4.1 and 4.2 enable us to study quantitatively how choice of the basis affects the efficiency of REACT estimators. To formulate the notion of economy, consider for every $b \in [0,1]$, every $r>0$, and every $\sigma^2>0$ the ball

$$B(r,b,\sigma^2) = \{\xi\colon \mathrm{ave}(\xi^2) \le \sigma^2 r \text{ and } \xi_i = 0 \text{ for } i > bp\}. \tag{44}$$

Evidently, $B(r,b,\sigma^2)$ is a special case of the ellipsoid $E(r,a,\sigma^2)$ that arises when $a_i = 1$ for $1 \le i \le bp$ and $a_i = 0$ for $bp < i \le p$. A basis U for the linear model is clearly economical if, in the resulting canonical model, $\xi \in B(r,b,\sigma^2)$ for some small value of b and some value of $r>0$. While this formulation is too stringent to serve as a complete definition of economy, it yields an illuminating first result on the interplay between basis economy and the risk of REACT estimators.

Theorem 4. Suppose that $\hat{\sigma}^2$ satisfies (40). For every $r>0$, $b\in[0,1]$, and $\sigma^2>0$, the following two limits hold:

$$\lim_{p\to\infty} \sup_{\xi\in B(r,b,\sigma^2)} R_p(\hat{\xi}_{NS},\xi,\sigma^2) = \sigma^2 \min\{r,b\} > \sigma^2 rb/(r+b) \tag{45}$$

$$\lim_{p\to\infty} \sup_{\xi\in B(r,b,\sigma^2)} R_p(\hat{\xi}_{M},\xi,\sigma^2) = \sigma^2 rb/(r+b). \tag{46}$$

The asymptotic minimax risk over all estimators is

$$\lim_{p\to\infty} \inf_{\hat{\xi}} \sup_{\xi\in B(r,b,\sigma^2)} R_p(\hat{\xi},\xi,\sigma^2) = \sigma^2 rb/(r+b). \tag{47}$$

Limit (47) is the specialization of (37) when $a_i = 1$ for $1 \le i \le bp$ and $=0$ otherwise. In this case, it follows from (33) and (32) that $\lim_{p\to\infty}\nu_p(r,a,\sigma^2) = \sigma^2 rb/(r+b)$. Moreover, g_0 has coefficients $g_{0,i} = [1-\mu^{-1/2}]_+$ for $1 \le i \le bp$ and $=0$ otherwise. Because $g_0 \in \mathcal{F}_M$, the reasoning in the last paragraph of Section 4.1 entails that $\tilde{f}_M z$ is asymptotically minimax over $B(r,b,\sigma^2)$ for every $r>0$, every $b\in[0,1]$, and every $\sigma^2>0$. This result together with Theorem 4.3 establishes (46).

Verification of (45) is by direct calculation of maximum risk. From (42) in Theorem 3,

$$\lim_{p\to\infty} \sup_{\mathrm{ave}(\xi^2)\le\sigma^2 r} R_p(\hat{\xi}_{NS},\xi,\sigma^2) = \lim_{p\to\infty} \sup_{\mathrm{ave}(\xi^2)\le\sigma^2 r} R_p(\tilde{\xi}_{NS},\xi,\sigma^2). \tag{48}$$

By the definitions of $\rho(f,\xi^2,\sigma^2)$ and of $\mathcal{F}_{NS}$,

$$R_p(\tilde{\xi}_{NS},\xi,\sigma^2) = \min_{f\in\mathcal{F}_{NS}} \mathrm{ave}[\sigma^2 f^2 + \xi^2(1-f)^2] = \min_{0\le k\le p} p^{-1}[\sigma^2 k + \sum_{i>k}\xi_i^2]. \tag{49}$$

Therefore, when $\xi \in B(r, b, \sigma^2)$,

$$R_p(\tilde{\xi}_{NS}, \xi, \sigma^2) \leq \min\{p^{-1} \sum_{i \leq bp} \xi_i^2, \sigma^2 b\} \leq \sigma^2 \min\{r, b\}. \tag{50}$$

On the other hand, let $\xi_i^2 = r\sigma^2/b$ for $1 \leq i \leq bp$ and $= 0$ otherwise. The vector ξ so defined clearly lies in $B(r, b, \sigma^2)$. Moreover, from (49),

$$R_p(\tilde{\xi}_{NS}, \xi, \sigma^2) = \min_{0 \leq k \leq p} p^{-1}(\sigma^2 k + [\lfloor bp \rfloor - k]_+ r\sigma^2/b) = \sigma^2 \min\{r, b\} + O(p^{-1}), \tag{51}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. Consequently,

$$\sup_{\xi \in B(r,a,\sigma^2)} R_p(\tilde{\xi}_{NS}, \xi, \sigma^2) \geq \sigma^2 \min\{r, b\} + O(p^{-1}). \tag{52}$$

Together, (50) and (52) establish (45).

The most important implications of Theorem 4 are as follows:

(a) The asymptotic minimax bound on the risk of any estimator over $B(r, b, \sigma^2)$ is achieved by $\hat{\xi}_M$. For fixed b and σ^2, this asymptotic minimax bound increases monotonically in r but never exceeds $\sigma^2 b$. The asymptotic maximum risk of $\hat{\xi}_{NS}$ also increases monotonically in r but never exceeds $\sigma^2 b$. However, the risk of the least squares estimator is σ^2 for every value of ξ. Thus, economy of the basis in the sense of small b makes the asymptotic maximum risk of REACT estimators far smaller than that of the least squares estimator. The estimated risks for $\hat{\eta}_M$ and $\hat{\eta}_{NS}$ in the data analyses of Sections 2.5 and 3.1 reflect this ability of REACT estimators to improve greatly on least squares.

(b) The asymptotic maximum risk of $\hat{\xi}_{NS}$ is never more than twice the asymptotic maximum risk of $\hat{\xi}_M$. The worst risk ratio of 2 occurs when $r = b$. In the data analyses cited above, the estimated risks were in fact close to one another. The reduction in risk achieved by a REACT estimator through using monotone shrinkage rather than nested selection comes distant second to the reduction achieved through using an economical basis for the regression space.

The monotone class $\mathcal{F}_M$ of shrinkage vectors is not the smallest generating an estimator of ξ that is asymptotically minimax over each $B(r, b, \sigma^2)$. An enrichment of $\mathcal{F}_{NS}$ suffices for that purpose. In the notation of Section 2.2, consider the nested selection with shrinkage class

$$\mathcal{F}_{NSS} = \bigcup_{0 \leq c \leq 1} \bigcup_{k=0}^{p} \{e(k)\}. \tag{53}$$

Let $\hat{f}_{NSS} = \text{argmin}_{f \in \mathcal{F}_{NSS}} \hat{\rho}(f)$ and $\hat{\xi}_{NSS} = \hat{f}_{NSS} z$. Inspection of the argument given above for (46) establishes that $\hat{\xi}_M$ may be replaced in (46) with $\hat{\xi}_{NSS}$. While interesting technically, this result should not be viewed as a recommendation to use $\hat{\xi}_{NSS}$ rather than $\hat{\xi}_M$. The latter estimator continues to behave well under the more general definition of economy that we consider next.

Let

$$\mathcal{E}_M(b) = \{a \in R^p : a_i = 1 \text{ for } 1 \le i \le bp, 1 \le a_{\lfloor bp \rfloor + 1} \le \ldots \le a_p \le \infty\}. \tag{54}$$

Less restrictively than above, we might say that a basis is economical if, in the resulting canonical model, $\xi \in E(r, a, \sigma^2)$ for some $r > 0$, some $a \in \mathcal{E}_M(b)$, and some small value of b. Evidently,

$$B(r, b, \sigma^2) \subset E(r, a, \sigma^2) \quad \text{if } a \in \mathcal{E}_M(b). \tag{55}$$

For every $r > 0$, every $a \in \mathcal{E}_M(b)$, every $b \in (0, 1]$, and every $\sigma^2 > 0$, it follows from (55), Theorem 1 and (47) that

$$\liminf_{p \to \infty} \nu_p(r, a, \sigma^2) \ge \sigma^2 r b/(r + b) > 0. \tag{56}$$

The shrinkage vector g_0, defined in (32), lies in $\mathcal{F}_M$ whenever $\xi \in E(r, a, \sigma^2)$ with $a \in \mathcal{E}_M(b)$. Consequently, the first part of Theorem 2 and the reasoning at the end of Section 4.1 yield the following result.

Theorem 5. Suppose that $\hat{\sigma}^2$ satisfies (40). Then, for every $r > 0$, every $a \in \mathcal{E}_M(b)$, every $b \in (0, 1]$, and every $\sigma^2 > 0$,

$$\lim_{p \to \infty} [\sup_{\xi \in E(r,a,\sigma^2)} R_p(\hat{\xi}_M, \xi, \sigma^2) / \inf_{\hat{\xi}} \sup_{\xi \in E(r,a,\sigma^2)} R_p(\hat{\xi}, \xi, \sigma^2)] = 1. \tag{57}$$

The monotone REACT estimator $\hat{\eta}_M$ is thus asymptotically minimax for any basis that is economical in the general sense stated after (54), in addition to being asymptotically minimax in the more restricted setting of Theorem 4.

4.4. Confidence sets

The following two theorems find the asymptotic distribution of $\hat{t}_{\mathcal{F}}$, defined in (25), and determine the asymptotic coverage probability and asymptotic loss of the confidence set $\hat{C}_{\mathcal{F}}$, defined in (27). Both results follow from Theorems 3.1 and 3.2 of Beran and Dümbgen (1999) upon recalling that $|\hat{\eta}_{\mathcal{F}} - \eta|^2 = |\hat{\xi}_{\mathcal{F}} - \xi|^2$. In the sequel, let d be any metric for weak convergence of probability measures on the real line and let $\mathcal{L}(\hat{t}_{\mathcal{F}})$ denote the distribution of $\hat{t}_{\mathcal{F}}$ under the model.

Theorem 6. Let $m = \min\{p, n - p\}$. Suppose that $\hat{\sigma}^2 = \hat{\sigma}^2_{LS}$, $\mathcal{F}$ is $\mathcal{F}_M$ or $\mathcal{F}_{NS}$, and

$$\lim_{m \to \infty} p/(n - p) = \gamma^2 < \infty. \tag{58}$$

Then, for every $r > 0$ and every $\sigma^2 > 0$,

$$\lim_{p \to \infty} \sup_{\text{ave}(\xi^2) \le \sigma^2 r} d[\mathcal{L}(\hat{t}_{\mathcal{F}}), N(0, \tau^2_{\mathcal{F}})] = 0, \tag{59}$$

where

$$\tau^2_{\mathcal{F}} = 2\sigma^4 \text{ave}[(2\tilde{f}_{\mathcal{F}} - 1)^2] + 2\gamma^2\sigma^4[\text{ave}(2\tilde{f}_{\mathcal{F}} - 1)]^2 + 4\sigma^2 \text{ave}[\xi^2(1 - \tilde{f}_{\mathcal{F}})^2]. \tag{60}$$

The variance $\tau_{\mathcal{F}}^2$ is a function of p, $n-p$, ξ^2, and σ^2. The estimator $\hat{\tau}_{\mathcal{F}}^2$ defined in (26) is obtained by substituting $z^2 - \hat{\sigma}_{LS}^2$ for ξ^2, $\hat{f}_{\mathcal{F}}$ for $\tilde{f}_{\mathcal{F}}$, and $\hat{\sigma}_{LS}^2$ for σ^2 in (60). Theorem 7 below establishes the consistency of $\hat{\tau}_{\mathcal{F}}^2$ in the sense that $\hat{\tau}_{\mathcal{F}}^2 - \tau_{\mathcal{F}}^2$ converges in probability to zero. Because the α-th quantile of the $N(0, \tau_{\mathcal{F}}^2)$ distribution is estimable by $\hat{\tau}_{\mathcal{F}} \Phi^{-1}(\alpha)$, limit (59) leads to the confidence ball $\hat{C}_{\mathcal{F}}$ for η defined in (27).

Let $\hat{r}_{\mathcal{F}} = \hat{\rho}(\hat{f}_{\mathcal{F}}) + p^{-1/2} \hat{\tau}_{\mathcal{F}} \Phi^{-1}(\alpha)$. The performance of $\hat{C}_{\mathcal{F}}$ will be measured through its coverage probability $\mathrm{P}(\eta \in \hat{C}_{\mathcal{F}}$ and through its geometrical loss

$$L(\hat{C}_{\mathcal{F}}, \eta) = \sup_{\theta \in \hat{C}_{\mathcal{F}}} L(\theta, \eta) = [|\hat{\eta}_{\mathcal{F}} - \eta| + \hat{r}_{\mathcal{F}}]^2, \tag{61}$$

which treats $\hat{C}_{\mathcal{F}}$ as a set-valued estimator of η.

Theorem 7. Under the hypotheses for Theorem 3.3, for every $r > 0$ and every $\sigma^2 > 0$,

$$\begin{aligned} \lim_{p \to \infty, K \to \infty} \sup_{\mathrm{ave}(\xi^2) \le \sigma^2 r} \mathrm{P}[|L(\hat{C}_{\mathcal{F}}, \eta) - 4\rho(\tilde{f}_{\mathcal{F}}, \xi^2, \sigma^2)| \ge K p^{-1/2}] = 0 \\ \lim_{p \to \infty, K \to \infty} \sup_{\mathrm{ave}(\xi^2) \le \sigma^2 r} \mathrm{P}[|\hat{r}_{\mathcal{F}}^2 - \rho(\tilde{f}_{\mathcal{F}}, \xi^2, \sigma^2)| \ge K p^{-1/2}] = 0. \end{aligned} \tag{62}$$

For every $\epsilon > 0$,

$$\lim_{p \to \infty} \mathrm{P}[|\hat{\tau}_{\mathcal{F}}^2 - \tau_{\mathcal{F}}^2| > \epsilon] = 0. \tag{63}$$

Moreover,

$$\liminf_{p \to \infty} \inf_{\mathrm{ave}(\xi^2) \le \sigma^2 r} \tau_{\mathcal{F}}^2 > 0 \tag{64}$$

and

$$\lim_{p \to \infty} \sup_{\mathrm{ave}(\xi^2) \le \sigma^2 r} |\mathrm{P}(\eta \in \hat{C}_{\mathcal{F}}) - \alpha| = 0. \tag{65}$$

In particular, the classical least squares confidence set $\hat{C}_{LS}$, the nested selection confidence set $\hat{C}_{NS}$ and the monotone confidence set $\hat{C}_M$ each have asymptotic coverage probability α. The asymptotic geometrical loss of each confidence set is just four times the asymptotic risk of the estimator at the center of the confidence set. Thus, for every nontrivial coverage probability α, $\hat{C}_{LS}$ is no smaller asymptotically than either $\hat{C}_{NS}$ or $\hat{C}_M$. It is not too surprising that the efficiency of the estimator at the center of the confidence set should influence its geometrical size. However it is remarkable that this reduction in the size of the confidence set can be very substantial, as indicated by the estimated risks in Sections 2.5 and 3.1.

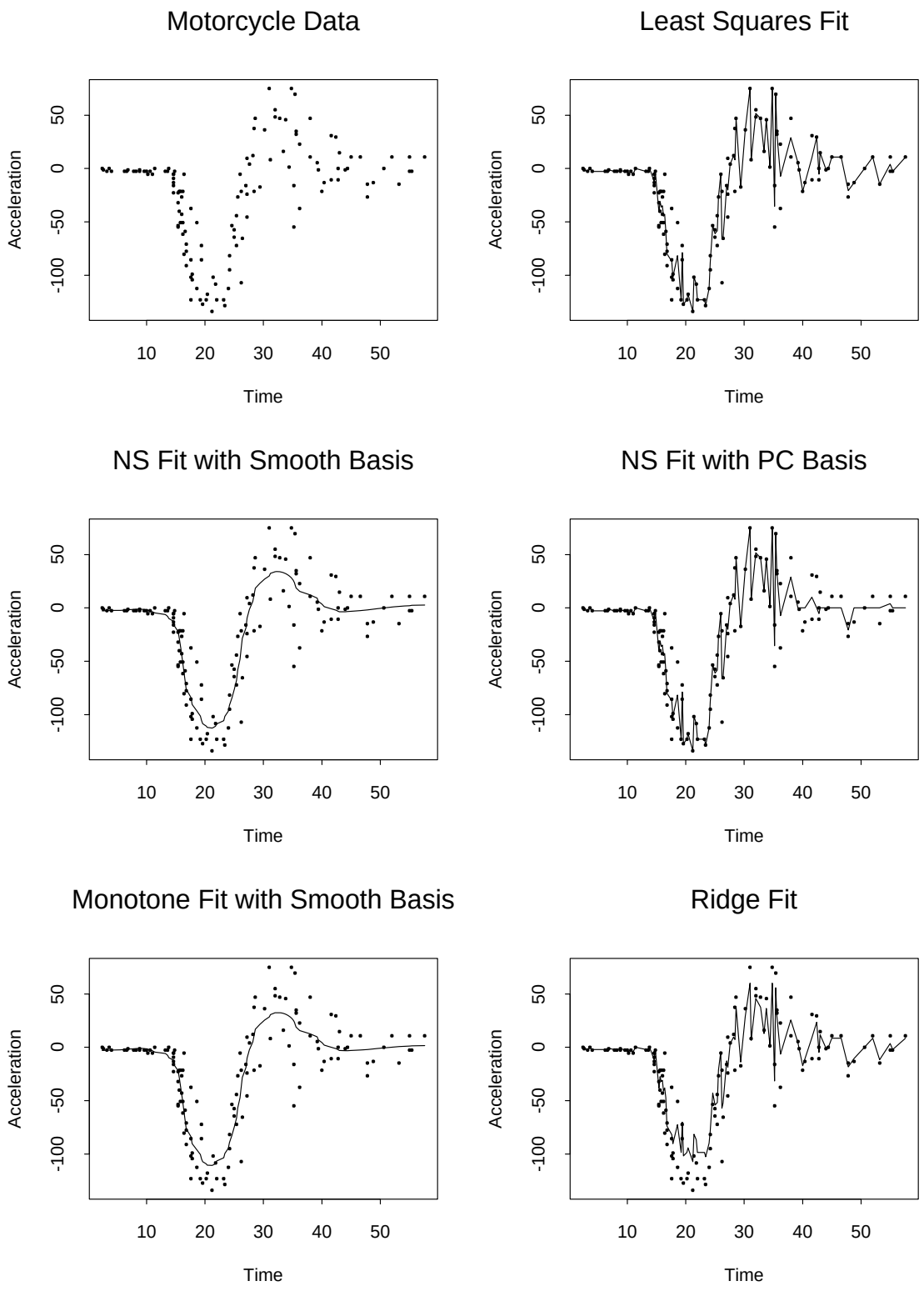

FIG 1. The left column displays the motorcycle data and the nested selection and monotone REACT fits using smooth basis U_S. The right column displays the least squares fit, the nested principal component fit, and the ridge fit.

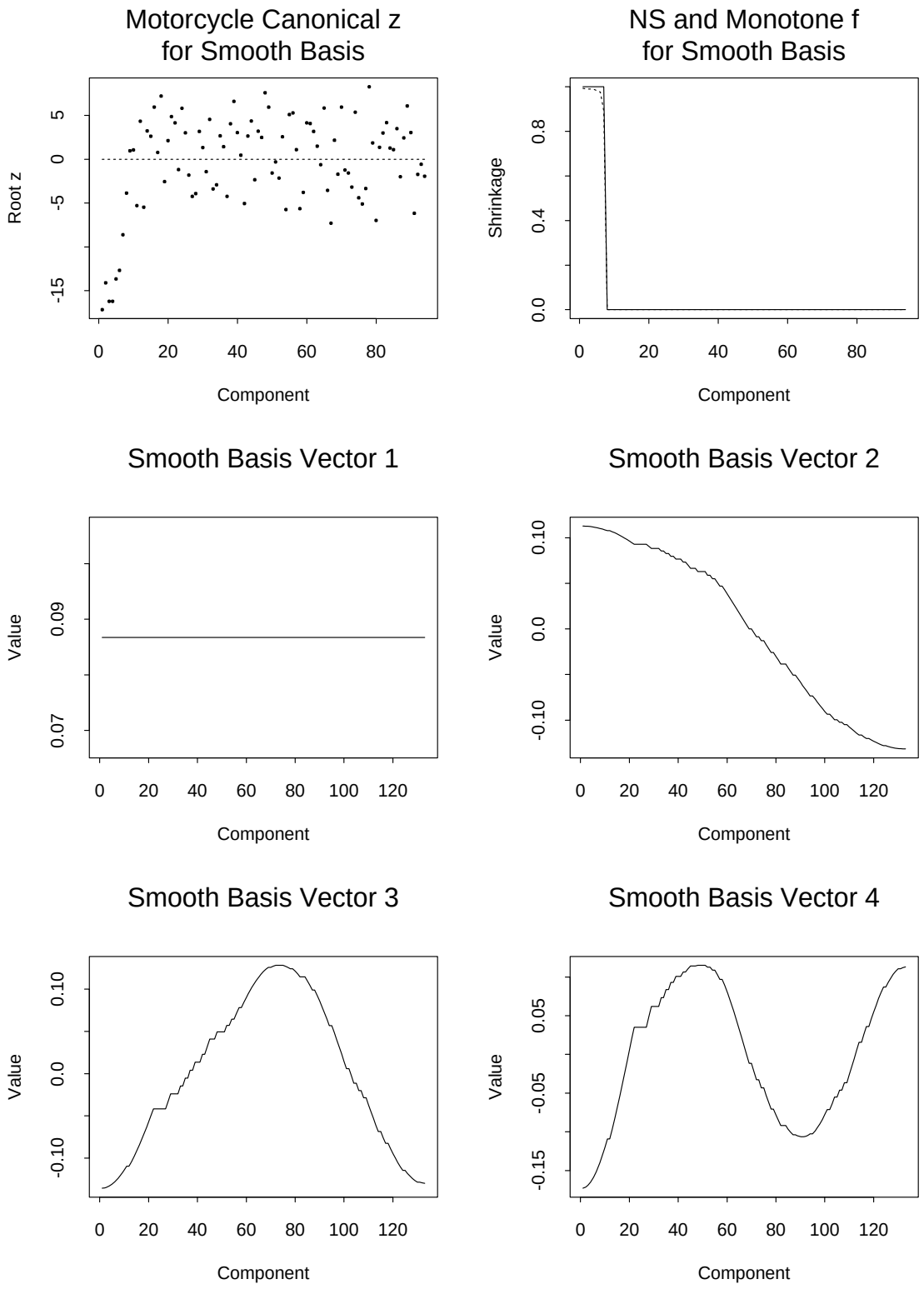

FIG 2. The second and third rows exhibit the first four vectors in the smooth basis U_S for the motorcycle data. The top row displays the canonical response $z = U_S'y$ and, on the right, the shrinkage vectors $\hat{f}_{NS}$ (solid line) and $\hat{f}_M$ (dashed line) for the smooth basis.

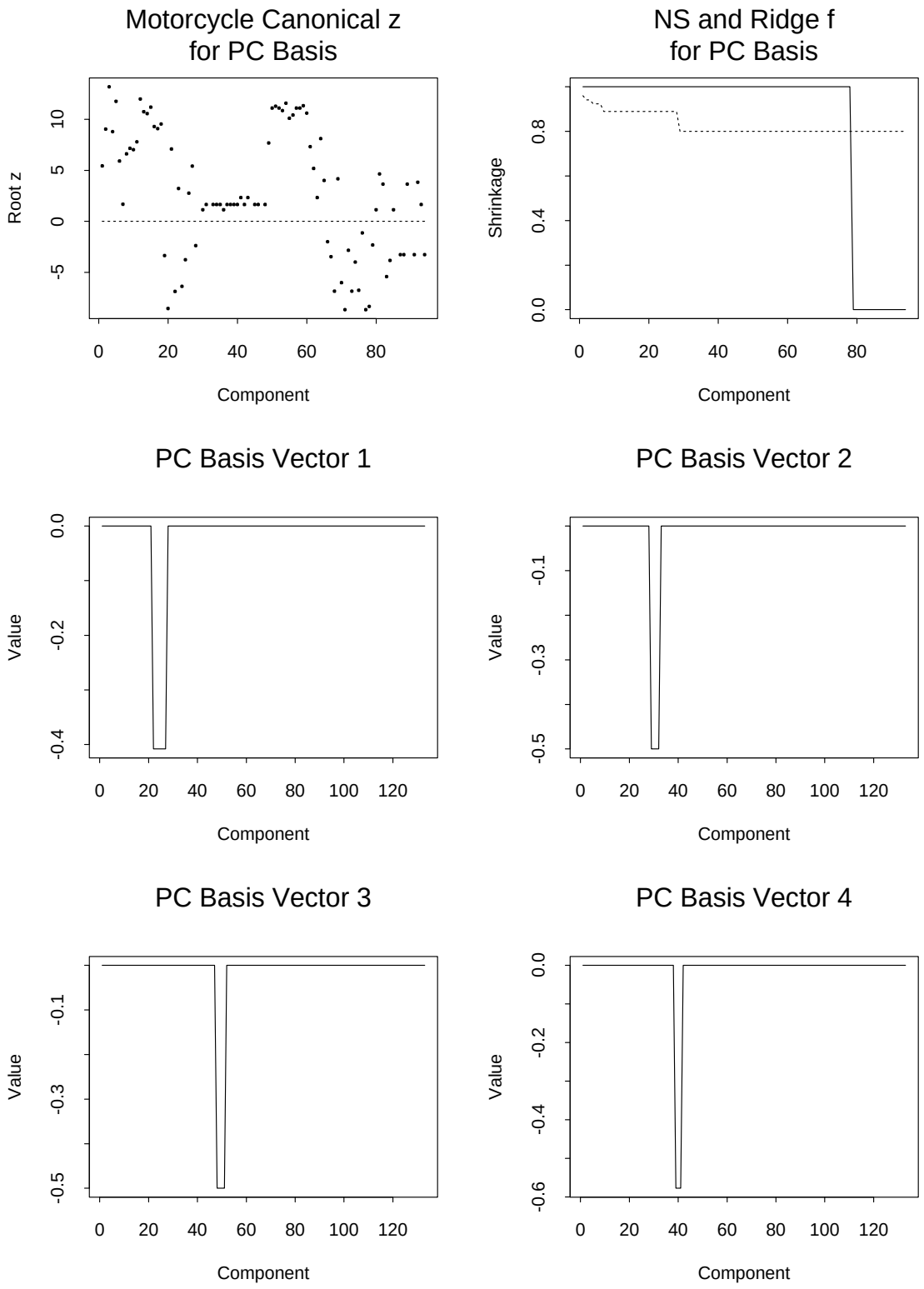

FIG 3. The second and third rows exhibit the first four vectors of the principal component basis U_{PC} for the motorcycle data. The top row displays the canonical response $z = U'_{PC}y$ and, on the right, the shrinkage vectors $\hat{f}_{PC}$ (solid line) and $\hat{f}_{RIDGE}$ (dashed line) for the PC basis.

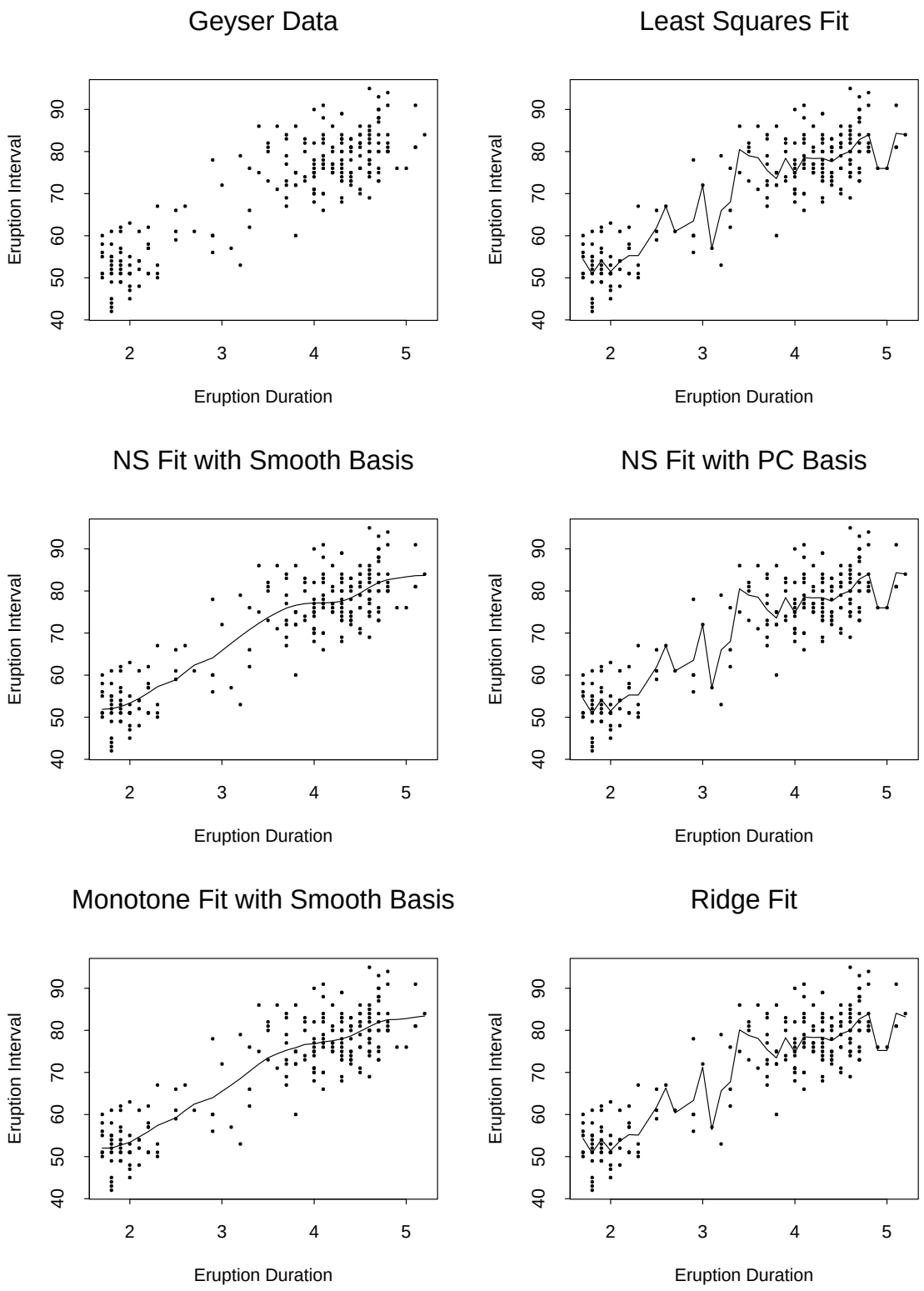

FIG 4. The left column displays the geyser data and the nested selection 4 and monotone REACT fits using smooth basis U_S. The right column displays the least squares fit, the nested principal component fit, and the ridge fit.

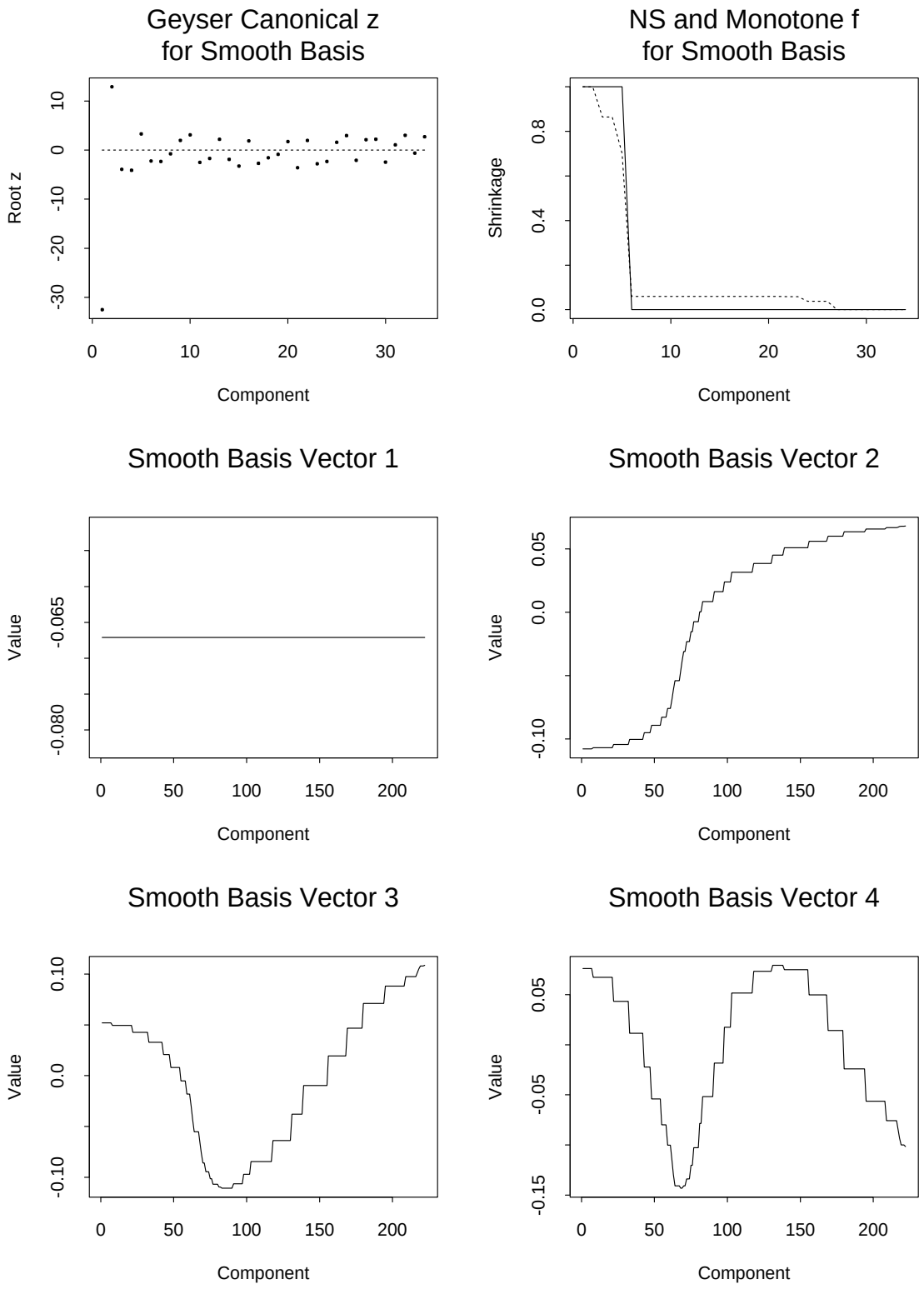

FIG 5. The second and third rows exhibit the first four vectors in the smooth basis U_S for the geyser data. The top row displays the canonical response $z = U_S'y$ and, on the right, the shrinkage vectors $\hat{f}_{NS}$ (solid line) and $\hat{f}_M$ (dashed line) for the smooth basis.

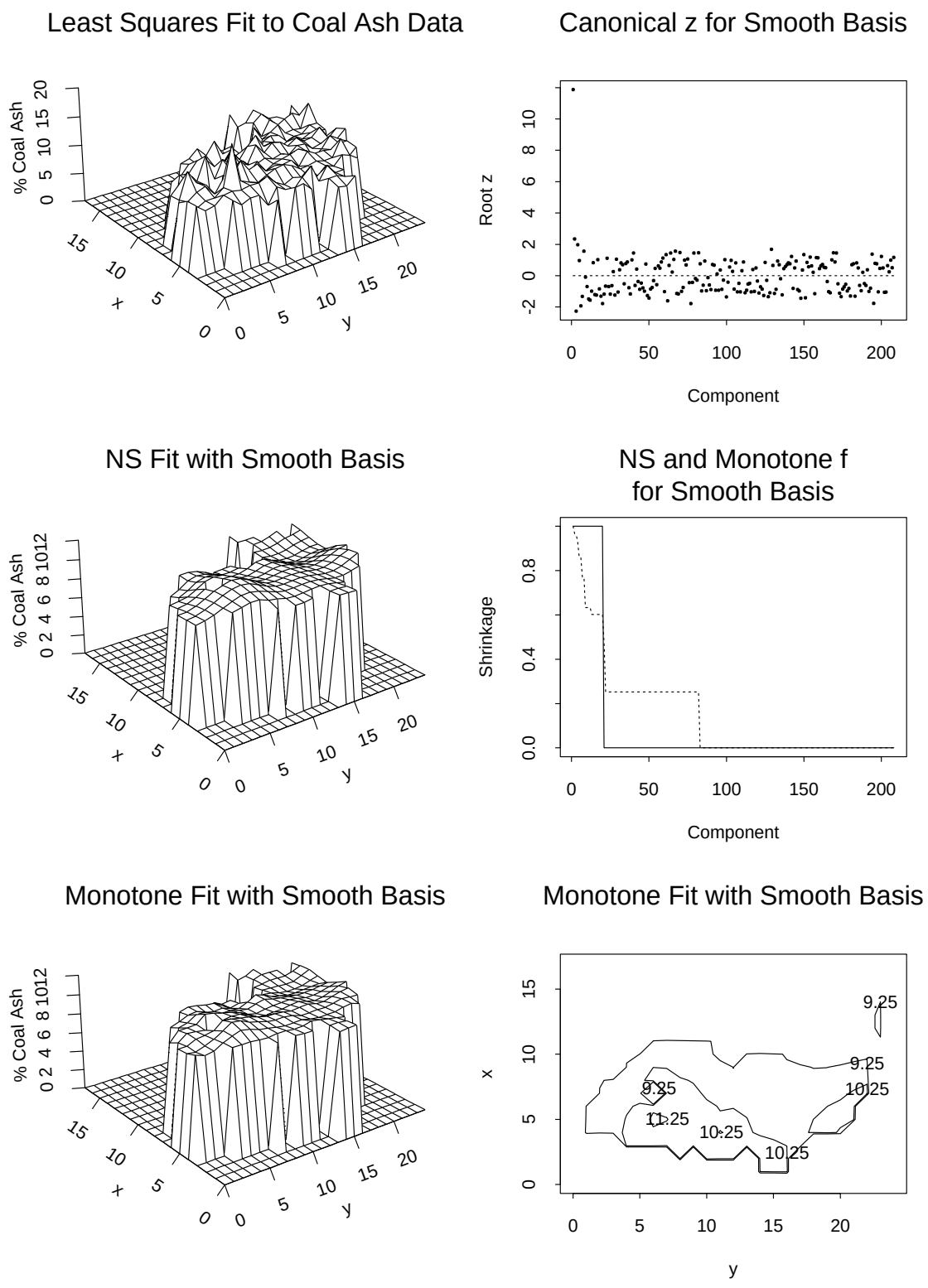

FIG 6. The left column displays the least squares fit and the nested selection and monotone REACT fits to the coal ash data using smooth basis U_{SS}. The right column displays the canonical response z, the shrinkage vectors $\hat{f}_{NS}$ (solid line) and $\hat{f}_M$ (dashed line), and a contour plot of the monotone smooth-basis fit.

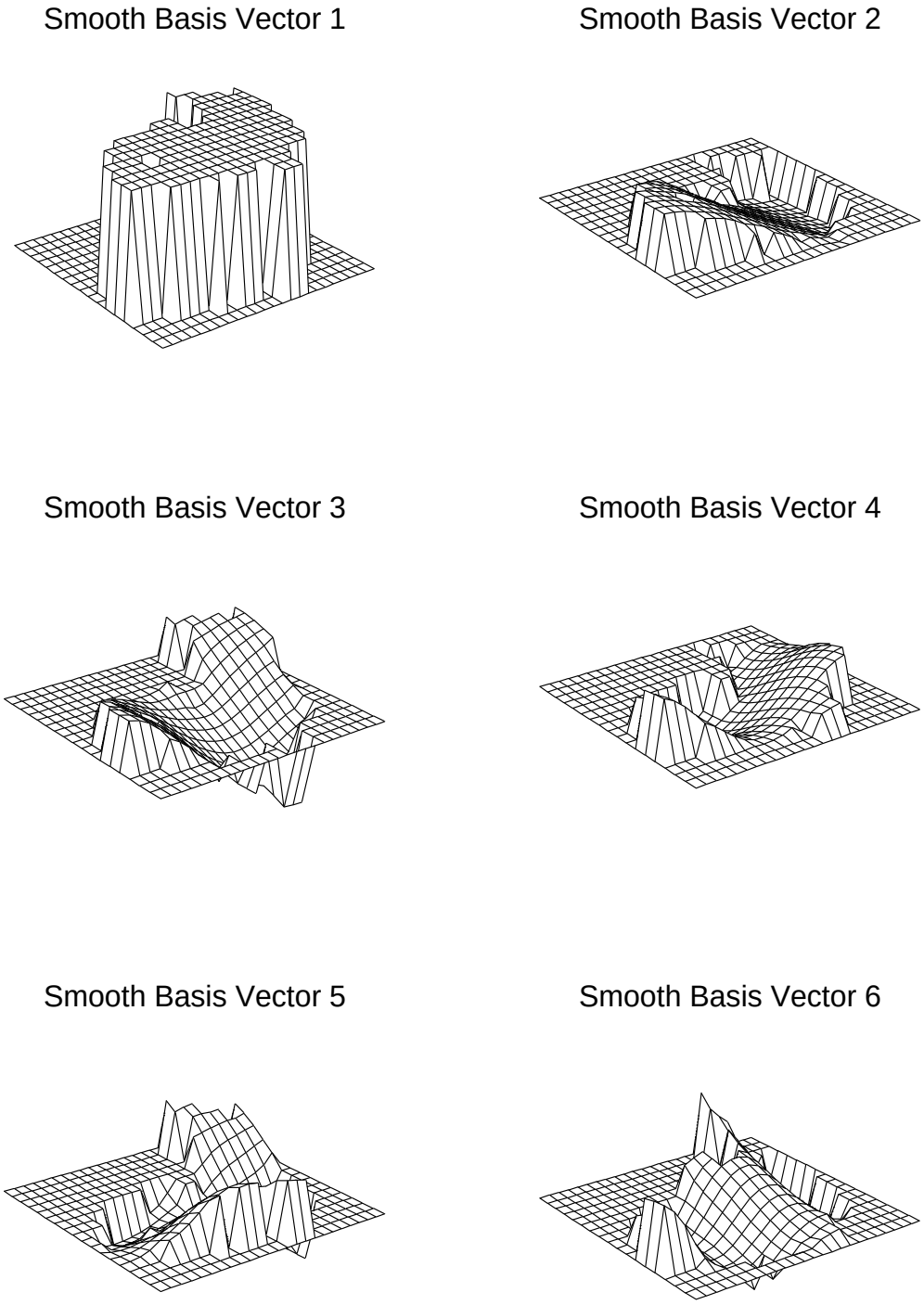

FIG 7. The first six vectors in the smooth basis U_{SS} for the coal ash data.

FIG 8. On the left is the smooth-basis monotone fit to the motorcycle data while on the right, spliced together, are separate smooth-basis monotone fits to the first 21 and remaining 112 data points. The separate REACT fits handle heteroscedasticity.

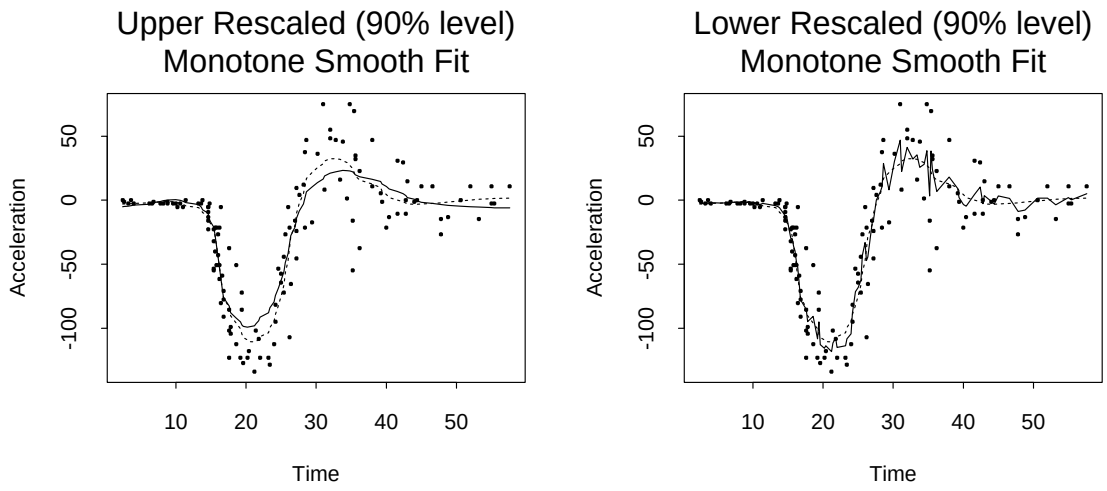

FIG 9. Displayed here (solid line) are the two smooth-basis monotone REACT fits, obtained by scaling $\hat{\sigma}^2$ upwards or downwards, that just lie on the boundary of the 90% confidence set. The dashed line is the monotone fit at the center of the confidence set.

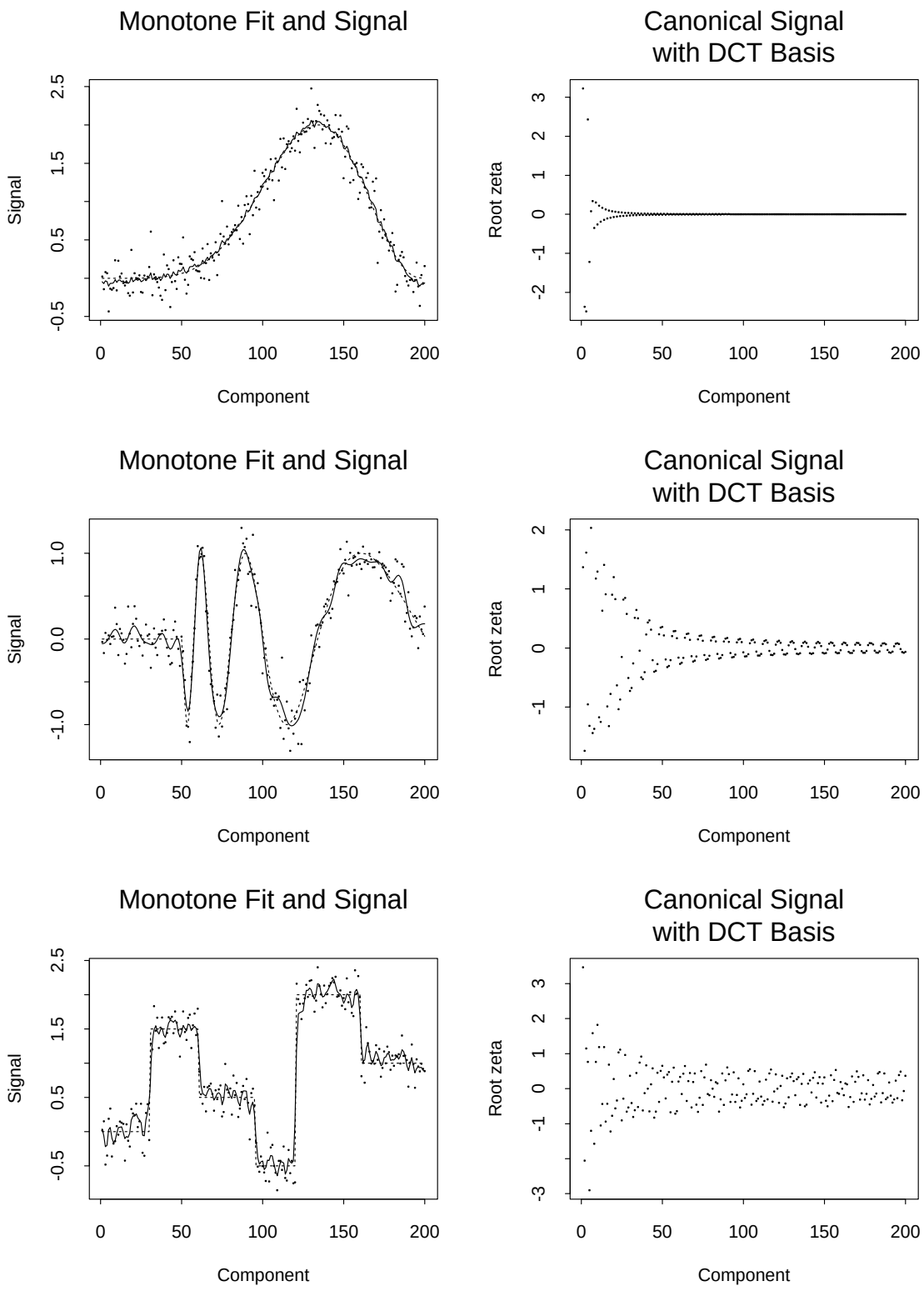

FIG 10. For three trend vectors η (dashed line), the left column displays a trend plus noise sample y of size 200 and its monotone REACT fit (solid line) using the discrete cosine basis U_{DC}. The canonical mean vectors $\xi = U'_{DC}\eta$ in the right column show that this basis is less economical for the second and third trends.

REFERENCES

Beran, R. (1996), "Confidence Sets Centered at C_p Estimators," *Annals of the Institute of Statistical Mathematics*, 48, 1–15.

Beran, R., and Dümbgen, L. (1999), "Modulation of Estimators and Confidence Sets," *Annals of Statistics*, in press.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models" (with discussion), *Annals of Statistics*, 17, 453–555.

Cressie, N. A. (1993), *Statistics for Spatial Data* (revised ed.), New York: Wiley.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

Efroimovich, S. Yu., and Pinsker, M. S., (1984), "Learning Algorithm for Nonparametric Filtering," *Automation and Remote Control*, 45, 1434–1440.

Fan, J., and Müller, M. (1995), "Density and Regression Smoothing," in *Xplore: An Interactive Statistical Computing Environment*, eds. W. Härdle, S. Klinke and B. A. Turlach, New York: Springer, 77–99.

Golubev, G. K. (1987), "Adaptive Asymptotically Minimax Estimators of Smooth Signals," *Problems of Information Transmission*, 23, 47–55.

Kneip, A. (1994), "Ordered Linear Smoothers," *Annals of Statistics*, 22, 835–866.

Li, K.-C. and Hwang, J. T. (1984), "The Data-smoothing Aspect of Stein Estimates," *Annals of Statistics*, 12, 887–897.

Li, K.-C. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation," *Annals of Statistics*, 13, 1352–1377.

Li, K.-C. (1987), "Asymptotic Optimality for C_p, C_L and Generalized Cross-validation: Discrete Index Set," *Annals of Statistics*, 15, 958–976.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–676.

Mallows, C. L. (1980), "Some Theory of Nonlinear Smoothers," *Annals of Statistics*, 8, 695–715.

Pinsker, M. S. (1980), "Optimal Filtration of Square-integrable Signals in Gaussian Noise," *Problems of Information Transmission*, 16, 120–133.

Rao, C. R., and Toutenberg, H. (1995), *Linear Models. Least Squares and Alternatives*, New York: Springer.

Rao, K. R., and Yip, P. (1990), *Discrete Cosine Transform. Algorithms, Advantages, Applications*, Boston: Academic Press.

Rice, J. (1984), "Bandwidth Choice for Nonparametric Regression," *Annals of Statistics*, 12, 1215–1230.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference*, New York: Wiley.

Silverman, B. W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting" (with discussion), *Journal of the Royal Statistical Society B*, 47, 1–52.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York: Springer.

Stein, C. (1956), "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, Berkeley: University of California Press, 197–206.

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *Annals of Statistics*, 9, 1135–1151.

Venables, W. N., and Ripley, B. D. (1997), *Modern Applied Statistics with S-PLUS* (second ed.), New York: Springer.